Standzeitenanalyse von Elektrorollern im Free-Floating-Sharing-System in Berlin

Non-deployment analysis of electric scooters in a free-floating-sharing system in Berlin

vorgelegt von

Mareike Vogel
Matrikel-Nummer: 569202

Bachelorarbeit eingereicht bei:

Erstprüfer:	Prof. Dr. W. Härdle
Zweitprüfer:	Prof. Dr. B. López Cabrera
Betreuer:	Dr. S. Klinke

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.) in Betriebswirtschaftslehre

Humboldt-Universität zu Berlin
Wirtschaftswissenschaftliche Fakultät
Institut für Statistik und Ökonometrie
Ladislaus von Bortkiewicz Lehrstuhl für Statistik

Berlin, den 16.02.2018

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

FFSS	Free Floating Sharing System
LOR	Lebensweltlich Orientierte Räume
ÖPNV	Öffentlicher Personennahverkehr

1. Einleitung

Nach einer langsamen Entwicklung des Rollersharingmarktes im Jahr 2015, erfährt der Markt in den Folgejahren 2016 und 2017 ein starkes Wachstum. Derzeit bieten weltweit 30 Städte ein Rollersharingsystem an, fast 80% dieser Städte sind europäisch. Allein 41% der Roller im Sharingangebot befinden sich in den europäischen Städten Berlin und Paris (Innoz, 2017). Diese Zahlen über den weltweiten Rollersharingmarkt veröffentlichte das Innovationszentrum für Mobilität und gesellschaftlichen Wandel im Jahr 2017. Im Gegensatz zum Car- oder Bikesharing, welche schon seit 1997 zunächst in Deutschland und der Schweiz, später auch weltweit angeboten werden, konzentriert sich der Rollersharingmarkt hauptsächlich auf stationsungebundene Systeme. (Bundesverband CarSharing, 2017). Anders als beim stationsgebundenen Sharing können beim stationsungebundenen System die Fahrzeuge unabhängig von einzelnen Stationen innerhalb eines bestimmten Geschäftsgebietes abgestellt werden. Diese Flexibilität des Anmiet- und Abstellortes hat zur Folge, dass die flächendeckende Verteilung der Fahrzeuge nicht mehr vom Betreiber, sondern von den Sharingnutzern abhängig ist.

Ein in der Wissenschaft bereits adressiertes Problem, dass sich aus dieser Art der Fahrzeugflottenführung herausbildet, ist die Relokalisierung der Fahrzeuge. Es stellt sich die Frage, wie man die Fahrzeuge flächendeckend und der Nachfrage entsprechend bestmöglich im Geschäftsgebiet verteilt bzw. den Kunden einen Anreiz schafft, die Fahrzeuge möglichst dort abzustellen, wo sie in kürzester Zeit erneut nachgefragt werden. Um einen Impuls setzen zu können, müssen die hoch- und geringfrequentierten Gebiete eines Geschäftsgebietes herausgearbeitet werden und die Absichten der Kunden, weshalb ein Fahrzeug zu welcher Zeit und an welchem Ort gemietet wird, nachvollziehbar gemacht werden. Mit der Erforschung des Mobilitätsverhaltens der Kunden in Sharingkonzepten beschäftigen sich sowohl Wissenschaftler im Car- und Bikesharing- als auch im Rollersharingmarkt. Anhand eines Datensatzes der Firma Electric Mobility Concepts GmbH (Im Folgenden „emmy“ genannt) aus Berlin wird in dieser Arbeit die Nachfrage nach Rollern im stationsungebundenen Sharing untersucht. Dazu werden die Stand- bzw. Wartezeiten der Roller in einem Zeitraum von sechs Wochen ermittelt und analysiert. Mithilfe von verschiedenen statistischen Regressionsmodellen soll außerdem beantwortet werden, welche Faktoren einen Einfluss auf die Standzeit der Fahrzeuge und damit der Nachfrage haben und in welchem Ausmaß.

Im Folgenden wird zunächst ein Überblick über die Begrifflichkeiten im Sharingmarkt gegeben und das Geschäftsmodell des Rollersharinganbieters emmy vorgestellt. Außerdem wird ein allgemeiner Überblick über den Stand der Forschung, die Relevanz der bisherigen Forschungsarbeiten und die Forschungsfrage dieser Arbeit erläutert. Im dritten Kapitel wird ein Überblick über den zu analysierenden Datensatz der Firma emmy gegeben. Nachdem die Datenbereinigung beleuchtet wird, folgt zunächst eine Übersicht zu allen verwendeten Datensätzen und –quellen und letztendlich die explorative Statistik des Hauptdatensatzes. Die statistische Methode wird im vierten Kapitel zweigeteilt anhand der Methode der linearen Regression und der Ereigniszeitenanalyse anhand der Cox-Regression beschrieben. Anwendung finden die beiden Methoden im fünften Kapitel, in dem die Ergebnisse der statistischen Auswertung über die Standzeitenanalyse interpretiert werden. Im letzten Kapitel werden die Ergebnisse zusammengefasst und ein Ausblick über mögliche weitere Ansätze in der Forschung von stationsungebundenem Sharing gegeben.
Die Datenanalyse wurde unter Verwendung der Statistiksoftware SPSS und des Geoinformationssystem QGIS durchgeführt.

2. Rollersharing und aktuelle Herausforderungen

In diesem Kapitel wird zunächst ein allgemeiner Überblick über den weltweiten Sharingmarkt geschaffen und das Geschäftsmodell des Rollersharing-Anbieters emmy vorgestellt. Des Weiteren wird der Stand der Forschung zum Thema Shared Mobility beleuchtet und letztendlich die Forschungsfrage und die Zielsetzung dieser Arbeit konkretisiert.

2.1 Stationsgebundene und -ungebundene Sharingsysteme

Neben den Plattformanbietern, zu denen beispielsweise das klassische Angebot von Taxiunternehmern zählt, gibt es im Bereich der Shared Mobility weitere Anbieter, die Mobilitätsdienstleistungen mit der Bereitstellung ihrer eigenen Flotte verbinden. Der größte Markt für Sharingangebote ist der Carsharingmarkt, gefolgt vom Bike- und Rollersharingmarkt. Definiert wird der Carsharingmarkt als eine organisierte, gemeinschaftliche Nutzung von Kraftfahrzeugen (Bundesverband Carsharing, 2017). Diese Definition ist auf den Rollersharingmarkt übertragbar, mit dem sich im Kontext dieser Arbeit ausschließlich befasst werden soll.

Der Sharingmarkt wird von Deutschland mit einem Angebot von insgesamt 2.495 Rollern angeführt, gefolgt von Frankreich an zweiter und Spanien an dritter Stelle (Innoz, 2017). Die Sharingangebote lassen sich in zwei Klassen einteilen: das stationsungebundene oder Free-Floating-Sharing und das stationsgebundene Sharing. Beim Free-Floating Sharing gibt es keine festen Stationen, die Fahrzeuge dürfen daher vom Kunden überall innerhalb des Geschäftsgebietes abgestellt werden. Dadurch sind die Kunden weitaus flexibler, da eine Rundfahrt genauso möglich ist wie ein One-Way-Trip, bei dem der Anmietort nicht mit dem Abstellort des Fahrzeuges übereinstimmt (Kortum, 2012). Beim stationsgebundenen Sharing hingegen gibt es feste Anlaufpunkte. Das Mietfahrzeug muss zu einem der Abstellorte zurückgebracht werden (Kortum, 2012). Anders als beim Carsharing basieren ein Großteil der Rollersharinganbietern auf dem Prinzip des Free-Floating-Sharings. Im Carsharing bieten nur 4 der 154 Anbieter in Deutschland stationsungebundenenes Sharing an. (Bundesverband CarSharing, 2017), zu diesen gehören unter anderem Car2Go und DriveNow. Zu den stationsgebundenen Anbietern hingegen zählen unter anderem flinkster, stadtmobil oder Greenwheels (Carsharing Magazin, 2017). Auf dem Rollersharingmarkt gehört beispielsweise Econduce aus Mexiko-Stadt zu den stationsgebundenen Anbietern (Innoz, 2017). Zu den stationsungebundenen Anbietern zählt das junge Unternehmen emmy aus Berlin, auf welches im folgenden Abschnitt detailliert eingegangen wird.

2.2 Emmys Geschäftsmodell in Berlin

Einer der größten Rollersharinganbieter weltweit ist das Startup emmy aus Berlin. Im Folgenden wird ein Überblick über emmys Geschäfts- und Preismodell gegeben.

Seit 2015 ist emmy auf dem deutschen Rollersharingmarkt aktiv und bot zunächst 150 Elektroroller zur Miete in Berlin an. Seit 2017 verbreitete emmy sein Mobilitätsangebot auf andere deutsche Städte wie Hamburg, München und Düsseldorf. Zur Flotte zählen nun insgesamt ca.1.000 Elektroroller, in Berlin sind davon ca. 400 Elektroroller platziert. Wie in Abbildung 1 dargestellt umfasst das Berliner Geschäftsgebiet das komplette Gebiet innerhalb des S-Bahn Ringes mit einer Fläche von knapp 88km^2. Emmy hat laut unternehmenseigenen Angaben ca. 42.000 registrierte Kunden, von denen ein Großteil in Berlin wohnen und ihren Account vollständig aktiviert haben, dh. sie haben zusätzlich zur Registrierung ihren Führerschein verifiziert (Stand: Oktober 2017). Es können sich alle Volljährigen mit mindestens einem Führerschein der Klasse B registrieren. Der Anmietprozess selbst wird über eine mobile App geregelt, in der den Kunden alle verfügbaren Roller auf einer Karte abgebildet werden (s. Abbildung 1). Nach der Reservierung eines Rollers, öffnet der Kunde mithilfe der mobilen App die Helmbox des Rollers, in welcher sich die zwei Helme und bei einigen der Rollermodelle auch der Schlüssel befindet. Nach der Fahrt, die immer innerhalb des Geschäftsgebietes beendet werden muss, stellt der Kunde den Roller wieder ab und verschließt die Helmbox über die App.

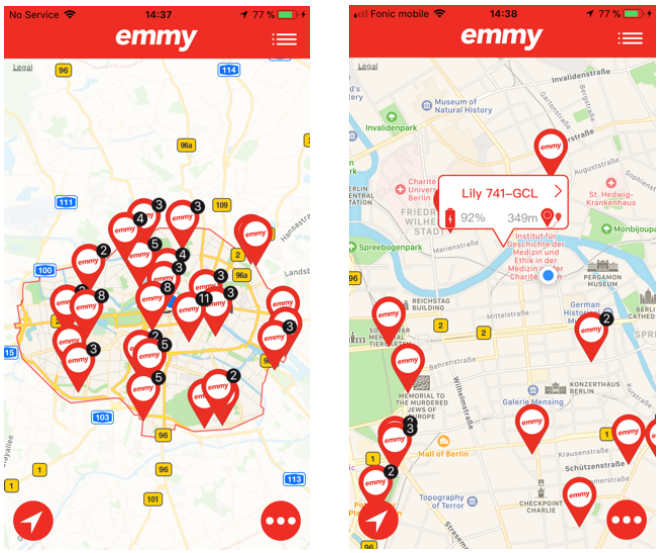

Abbildung 1: Emmys Geschäftsgebiet und Appansicht

Die Preisstruktur von emmy setzt sich aus einem Minuten- bzw. Kilometerpreis zusammen. Je nachdem, was für den Kunden günstiger ist, wird entweder 19ct/min oder 59ct/km am

Ende jeder Fahrt berechnet. Die einmalige Anmeldegebühr, die bei der Registrierung anfällt, beträgt 10€ inklusive 100 Freiminuten (emmy-sharing.de, 2017).
Die Kundenbasis des Startups und auch der anderen Rollersharinganbietern basiert zum Großteil aus „Young Urban Professionals“. Viele Kunden sind also zwischen 30 und 35 Jahren alt, besitzen kein eigenes Fahrzeug und sind bei mehr als nur einem der Sharinganbieter registriert (Innoz, 2017). Knapp 80% der emmy Kunden sind außerdem männlich. Da emmy die meisten Rollerfahrten in den Monaten zwischen Mai bis November verzeichnet, ist die Dienstleistung ein stark saisonaler Service. Im Winter behält sich das Startup vor, bei schlechten Wetterbedingungen, wie Schnee und Unwetter, den Service einzustellen.

2.3 Stand der Forschung

Die bisherige Forschung bietet aufgrund der Tatsache, dass Free-Floating-Sharing-Systeme (im Folgenden „FFSS“) erst seit 2011 auf dem Markt sind, kein weites Spektrum an Nutzerdatenauswertungen (Kortum et al, 2012). Die bisher veröffentlichte Literatur basiert zudem in großen Teilen auf Carsharingdaten, einige wenige Forschungen wurden auch zum Bikesharing durchgeführt. Sowohl Car-, Roller- als auch Bikesharing stellen sich weitestgehend ähnlichen Problemen zur Nutzungsoptimierung angesichts der Ungleichverteilung ihrer Flotte innerhalb des Geschäftsgebietes, jedoch lassen sich die Ergebnisse einer Carsharinguntersuchung nicht ohne weitere Annahmen und Analysen auf den Rollersharingmarkt übertragen. Viele Forschungsarbeiten zur Shared Mobility beschäftigen sich zusätzlich mit dem Thema, Strategien der Relokalisierung zu inszenieren und zu bewerten. Das Problem der Relokalisierung wird in dieser Arbeit jedoch nicht berücksichtigt.
Im Folgenden sollen zwei wichtige Veröffentlichungen von Schmöller und Bogenberger in Zusammenarbeit mit weiteren Autoren und die Forschungsarbeit von Kortum und Machemehl vorgestellt werden. Des Weiteren folgt eine Einschätzung, in wie weit die Ansätze und Ergebnisse der vorgestellten Forschungen Gemeinsamkeiten und Unterschiede mit der Standzeitenanalyse der Roller aufweisen könnten.

Bereits im Jahr 2012 haben sich Kortum und Machermehl mit dem Thema des FFSS beschäftigt. In ihrer Arbeit „Innovations in Membership prediction, modes share, and vehicle allocation optimization methologies” stellen sie die Ergebnisse einer umfassenden Analyse von Carsharingdaten des Anbieters Car2Go in Austin, Texas, vor.

Eine umfassende explorative Analyse zeigt die zeitliche und örtliche Verteilung der Nachfrage nach Autos im FFSS. Außerdem befassen sich Kortem et al. mit der Modellierung der Datensätze, um Prognosen ableiten zu können und eine optimale Flottenverteilung durch Relokalisierung zu erzielen. Auch der Zusammenhang des Altersdurchschnitts und der Anzahl an Buchungen wird dargestellt. Die statistische Herangehensweise zur Modellermittlung von Nutzungsdaten im FFSS wird in der folgenden Standzeitenanalyse aufgegriffen. Die unterschiedlichen sozioökonomischen Strukturen von Berlin und Austin und der Fortschritt des FFSS Marktes seit 2012 fordert jedoch eine aktuellere Untersuchung des Marktes.

Die Autoren Schmöller und Bogenberger haben in Zusammenarbeit mit weiteren Autoren verschiedene Forschungsarbeiten sowohl zum Thema Carsharing, als auch Bikesharing veröffentlicht. Untersucht wurden neben FFSS auch stationsgebundene Carsharingsysteme. Die Forschungsarbeit „Analyzing External Factors on the Spatial and Temporal Demand of CarSharingSystems“ wurde 2014 von Schmöller und Bogenberger veröffentlicht, welche sich mit dem Vergleich eines stationsgebundenem Carsharinganbieters und einem FFSS beschäftigt. Eine weitere Arbeit mit dem Titel „Empirical analysis of free-floating carsharing usage: The Munich and Berlin case“ wurde 2015 von Schmöller, Weikl, Müller und Bogenberger publiziert und vergleicht wiederum verschiedene Aspekte und deren Einfluss im FFSS in München und Berlin.
In beiden Arbeiten werden die Datensätze, die die Buchungsdaten von unterschiedlichen Zeitperioden beinhalten, nach zeitlichen und räumlichen Verteilungen ausgewertet. Außerdem werden die nachfragestärksten und -schwächsten Zeiten und Gebiete in den Städten dargestellt. In einem letzten Schritt wird die Nachfrage in Verbindung mit externen Faktoren wie dem Wetter und soziodemographischen Daten untersucht.
Schmöller et al. untersuchen anhand der Buchungsdaten die tatsächliche Nachfrage nach Autos, wobei eine Standzeitenanalyse, wie sie in dieser Arbeit durchgeführt wird, ein anderen Modellansatz beinhaltet. Daher kann beispielsweise auch das Forschungsergebnis von Schmöller et al., dass das Wetter nur einen geringen Einfluss auf die Nachfrage nach Carsharing in München hat, nicht ohne Weiteres übertragen werden. Zusätzlich ist ein Rollerfahrer den Wetterbedingungen stärker ausgesetzt als ein Autofahrer.
Der Einfluss von sozioökonomischen Strukturen in einzelnen Gebieten Münchens und Berlins auf die Nachfrage wird im weiteren Verlauf dieser Arbeit aufgegriffen. Die Forschungsarbeit von Schmöller et al. aus dem Jahr 2015 zeigt auf, dass einzelne Faktoren wie beispielsweise die Anzahl an Gewerben in einem bestimmten Gebiet in den Städten

Berlin und München unterschiedlich starken Einfluss haben. Außerdem fokussieren sich Carsharing- und Rollersharinganbieter auf die gleiche sozioökonomische Zielgruppe, weswegen der Zusammenhang eines hohen Anteils an jungen Leuten in einem Wohngebiet mit einer hohen Anzahl an Buchungen in der folgenden Standzeitenanalyse ähnliche Ergebnisse liefern wird.

2.4 Forschungsfrage und Ziel der Arbeit

Der vorliegende Datensatz, der einen Zeitraum von 6 Wochen umfasst, soll in dieser Arbeit analysiert werden und eine Grundlage zur Erarbeitung eines Relokalisierungsmodells durch die Firma emmy bilden. Neben der zeitlichen und räumlichen Verteilung ist es daher außerdem wichtig, die externen Einflussfaktoren auf die Standzeit bzw. die Nachfrage nach den Rollern zu verstehen, welches im fünften Kapitel untersucht wird. Die Standzeit wird in Zusammenhang mit meteorologischen, sozioökonomischen, zeitlichen als auch flotten bzw. rollerspezifischen Faktoren gesetzt. Ziel dieser Arbeit ist es, anhand dieser Analyse ein Muster in den historischen Daten des Rollersharinganbieters emmy zu erkennen und so die Rollernachfrage in der Zukunft prognostizieren zu können.

Durch das Verstehen der Verhaltensmuster kann der Sharingdienstleister im nächsten Schritt seinen Service besser an die Kunden anpassen. Dies soll in Form eines Relokalisierungsmodells erfolgen, welches in dieser Arbeit jedoch nicht inhaltlich untersucht wird. Ziel eines Relokalisierungsmodells wäre es, die Standzeiten der Roller durch beispielsweise Incentivierung der Kunden zu verkürzen, welches wiederum drei Herausforderungen des Unternehmens bewältigt. Zum einen können durch kürzere Standzeiten der Roller Schäden an den Fahrzeugen vorgebeugt werden. Eine konstante Inanspruchnahme der Batterien ist wie bei kraftstoffverbrauchenden Fahrzeugen auch bei Elektrorollern vorteilhaft. Gleichzeitig wird die Kundenzufriedenheit durch eine bedarfsgerechte Verfügbarkeit der Roller gesteigert, da die Roller an den nachgefragten Orten zu den relevanten Zeiten vorzufinden sind. Die Reduktion der Standzeiten ist aus Sicht des Unternehmens außerdem zur Auslastungs- und Profitabilitätssteigerung ein wichtiges Anliegen.

Zur eigentlichen Strategie der Relokalisierung gibt es verschiedene Theorien, wie die Incentivierung der Kunden durch vergünstige Fahrten oder die Umverteilung der Roller im Geschäftsgebiet durch das Unternehmen. Da die Erarbeitung eines Relokalisierungsmodells durch emmy selbst stattfindet, soll das aufgestellte Modell zur Standzeitenanalyse in dieser Forschungsarbeit allgemein gestaltet werden und eine Grundlage zur weiteren Forschung bilden.

3. Datensatz

Der zu analysierende Datensatz wird von der Firma emmy bereitgestellt. Er umfasst die Grundgesamtheit aller Mieten, die innerhalb von sechs Wochen im Zeitraum vom 15.September bis zum 19.Oktober 2017 in Berlin stattgefunden haben. Anhand der einzeln aufgelisteten Mieten wurde manuell die Standzeit eines Rollers zwischen zwei Mieten ermittelt, welches als Variable *standzeit* die Grundlage der Analyse bildet. Die umfangreiche Bereinigung des Datensatzes hat sowohl technische als auch methodische Gründe. Die wichtigsten Datenbereinigungen werden im nächsten Abschnitt erläutert, gefolgt von einer ausführlichen Beschreibung aller zusätzlichen Datenquellen für die Aufstellung von Regressionsmodellen, sowie die zeitliche und räumliche Verteilung des Nutzerverhaltens.

3.1 Datenbereinigung

Der Datensatz umschließt unter anderem den Zeitraum des Sturmes „Xavier“ vom 04.10. bis 06.10.2017. Der Sturm macht sich im vorliegenden Datensatz vor allem in der Windstärke an den besagten Tagen bemerkbar. Die durchschnittliche Windstärke in dem Zeitraum von Mitte September bis Mitte Oktober beträgt 3,13 km/h, an den Tagen des Sturmes jedoch 7,9 km/h bis hin zu einem Höchstwert von 16,8 km/h. Die eigentliche Hypothese, dass starker Wind zu längeren Standzeiten der Roller führt, kann anhand der Datenerfassung in diesem Zeitraum nicht bestätigt werden, was jedoch dem Einfluss anderer Verkehrsmittel zugrunde liegt. Am 05.10.17 stellten sowohl Straßenbahnen, S- und U-Bahnen ihren Betrieb aufgrund von Störungen und der vorhergesagten Gefahrenlage in Berlin ein. Außerdem war das System von Car- und Rollersharinganbietern wie Car2Go, DriveNow und Coup zeitweise offline. Da emmy von allen Mobilitätsanbietern zuletzt seinen Service offline schaltete, lässt sich aus dem Datensatz erkennen, dass die Nachfrage nach den Rollern anstatt zu sinken signifikant stieg. Da es sich hier um eine Extremsituation handelt, in dem spezielle Einflussfaktoren zur Nachfragesteigerung führten, wird dieser Zeitraum aus dem allgemeinen Datensatz ausgeschlossen.

Des Weiteren werden Standzeiten, die nach Berechnung des Anomalieindexes und zusätzlicher einzelner Überprüfung unrealistisch hoch erscheinen, aus dem Datensatz entfernt. Die Berechnung des Anomalieindexes für jeden Datenstrang ist eine statistische Methode um Fälle eines Datensatzes zu erkennen, die verhältnismäßig stark von den anderen Fällen abweichen. Da die einzelnen Standzeiten der Roller manuell und nicht per System automatisch ermittelt werden, kommt es wiederholt zu falschen Berechnungen. Die Roller,

die entweder in den inaktiven Modus umschalten, da der Akku aufgeladen werden muss, oder aufgrund von Störungen in die Werkstatt geholt werden müssen, sind für den Kunden in der mobilen App nicht sichtbar und können daher nicht angemietet werden. Diese inaktiven Roller gehen erst nach erfolgreicher Aufladung oder Reparatur wieder online. Bei der Berechnung der Standzeiten kann es vorkommen, dass der Wechsel eines Rollers in den inaktiven Modus aus den Daten selbst nicht erkannt werden kann, weswegen diese Fälle letztendlich eine sehr langen Standzeit aufweisen. Sie führen zu starken Verzerrungen in der Untersuchung der Einflussfaktoren. Anhand der Berechnung des Anomalieindex werden solche besagten Ausreißer aus dem Datensatz erkannt und aus der Analyse ausgeschlossen.

3.2 Allgemeine Angaben zur Datengrundlage

In der folgenden Analyse werden neben den Reservierungsdaten auch weitere Datensätze unterschiedlicher Quellen berücksichtigt. Die Lebensweltlich Orientierten Räume der Stadtverwaltung Berlin dienen zur geographischen Einteilung und werden im Abschnitt 3.2.1 näher erläutert. Außerdem werden die Kundenwohnorte anhand eines zweiten Datensatzes von emmy bereitgestellt, auf welchen in Abschnitt 3.2.2 näher eingegangen wird. Neben der Stadtverwaltung Berlin wird außerdem der Deutsche Wetterdienst als externe Quelle verwendet, um die meteorologischen Einflussfaktoren auf die Standzeit zu untersuchen. In Abschnitt 3.2.3 wird dieser Datensatz vorgestellt.

3.2.1 Lebensweltlich Orientierte Räume Berlin

In der weiteren Arbeit wird die Stadt Berlin geographisch anhand der Lebensweltlich Orientierten Räumen Berlin (Im Folgenden „LOR“ genannt) betrachtet. Diese wurden 2006 in gemeinsamer Abstimmung zwischen den planenden Fachverwaltungen des Senats, den Bezirken und dem Amt für Statistik Berlin-Brandenburg auf Grundlage von bereits definierten Sozialräumen festgelegt. Die räumliche Einteilung der Stadt Berlin in LOR wurde anhand von Kriterien der Baustrukturen, großer Straßen, Verkehrstrassen und natürlichen Grenzen erarbeitet. Ziel bei der Definition der LOR war es, lebensweltliche Homogenität abzubilden und zudem Vergleichbarkeit der Räume zu gewährleisten. Insgesamt stellt die Stadt Berlin drei Ebenen auf ihrer Website als Download zur Verfügung, von welchem in dieser Arbeit die Einteilung in 447 Planungsräume Gebrauch gemacht wird (Stadtentwicklung.berlin.de, 2017). Die einzelnen Räume sind kleinteilig genug, um die unterschiedlichen Standzeiten im Geschäftsgebiet hinweg aussagekräftig abbilden zu können und an gleicher Stelle die sozioökonomischen Daten wie beispielsweise

Einwohnerzahl, Gewerbeanteil oder ÖPNV-Haltestellen der einzelnen LOR mit einfließen zu lassen. 158 dieser Planungsräume befinden sich dabei vollständig oder zumindest teilweise innerhalb des Geschäftsgebietes von emmy.

3.2.2 Allgemeine Angaben zur Kundenbasis

Emmy zählt rund 42.000 registrierte Kunden (Stand: Oktober 2017), die sich auf fünf deutsche Städte verteilen. Im Datensatz zur Kundenbasis werden diejenigen Kunden berücksichtigt, die ihren Wohnort in Berlin angegeben haben. Es kann davon ausgegangen werden, dass ein Kunde, der in Berlin wohnt, den Sharinganbieter öfter nutzt als ein Tourist oder Wochenendbesucher. Zunächst wurden die in Berlin lebenden Kunden aus dem Datensatz herausgefiltert und im nächsten Schritt die Adressen den einzelnen LOR in Berlin zugeordnet. Abbildung 2 zeigt die räumliche Verteilung der Kundenwohnorte in Berlin, wobei die Anzahl der registrierten Kunden ins Verhältnis zur jeweiligen Fläche des LOR in km^2 gesetzt wurde. Je dunkler ein LOR gefärbt ist, desto mehr Kunden pro Fläche wohnen in diesem LOR. Beliebte Wohngebiete für emmys Kunden sind klar an den Außengrenzen des Geschäftsgebietes zu erkennen, das Zentrum Berlins weist durch lediglich eine helle Färbung im Vergleich weniger Kunden pro km^2 auf. Außerdem wohnen im Vergleich wesentlich mehr Kunden im östlichen Teil der Stadt, wie beispielsweise in den Bezirken Prenzlauer Berg, Kreuzberg-Friedrichshain und Neukölln, als in den westlich liegenden Bezirken wie Schöneberg und Charlottenburg.

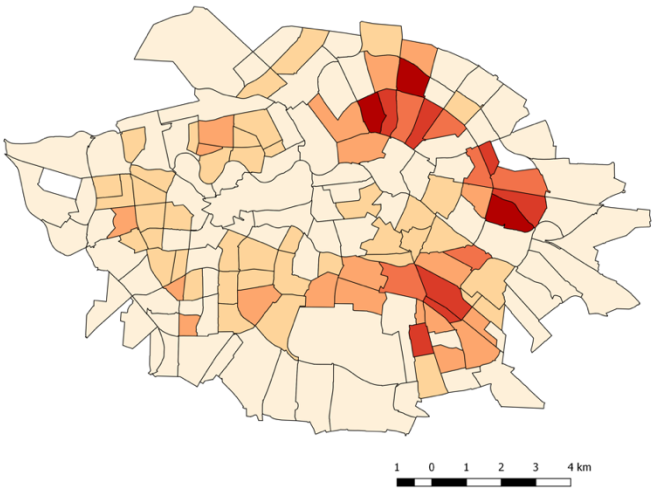

Abbildung 2: Registrierte Kunden / Fläche des LOR

Unabhängig von der Bezirksfläche ergibt sich aus dem Datensatz zu emmys Kundenbasis außerdem, dass die meisten Kunden in den Stadteilen Kreuzberg-Friedrichshain (ca. 24%), Berlin Mitte (ca. 19%), Prenzlauer Berg (ca. 16%) und Neukölln (ca. 11%) wohnen.

3.2.3 Wetterdaten und Verknüpfung mit den Nutzerdaten

Der Deutsche Wetterdienst erhebt stündlich verschiedene meteorologische Kennzahlen, welche über die Startzeitstunde jeder einzelnen Reservierung mit den Standzeitendaten verknüpft werden. Es werden Temperatur in °C, Niederschlag in mm/km^2 und die Windstärke in km/h in Betracht gezogen (Dwd.de, 2017). Unter der Annahme, dass sich diese drei Faktoren über Berlin hinweg nicht signifikant unterscheiden, werden alle Datenstränge unabhängig von ihrem genauen Standort mit den Erhebungen der Wetterstation Berlin Tempelhof verknüpft.

3.3 Emmys Nutzerdaten im Erhebungszeitraum

Nach der Datenbereinigung beinhaltet der Datensatz im Erhebungszeitraum von insgesamt 6 Wochen eine Grundgesamtheit von 23.800 Datenstränge zu Standzeiten. Im Folgenden werden die zeitliche und räumliche Verteilung der Nachfrage bzw. der Standzeiten der einzelnen Roller erläutert und ein Überblick über die einzelnen Variablen, die in Kapitel 5 als zu erklärende Variablen der Standzeit verwendet werden, geschafft.

3.3.1 Zeitlicher Verlauf der Standzeiten

Der zeitliche Verlauf der Rollernachfrage ist für jeden Wochentag einzeln in Abbildung 3 abgebildet. Alle Fahrten wurden anhand der Startzeit in Stundenfenster gruppiert. Die Stunde 11 beispielsweise umfasst alle Fahrten, die zwischen 11:00 und 11:59Uhr begonnen haben.

Für alle Werktage von Montag bis Freitag lässt sich ein ähnlicher Verlauf erkennen mit zwei Spitzen um 10-11 und 17-21 Uhr. An Samstagen und Sonntagen, in Abbildung 3 in blau markiert, hingegen lässt sich ein flacherer Anstieg der Nachfrage über den Vormittag hinweg erkennen bis hin zum Höhenpunkt zwischen 14 – 18 Uhr. In der folgenden Analyse wurden daher alle Fahrten nach Werktagen bzw. Wochenenden zusammengefasst. Aufgrund der unterschiedlichen zeitlichen als auch räumlichen Verteilung der Daten, wird der Datensatz im späteren Verlauf dieser Analyse in zwei Teile gegliedert.

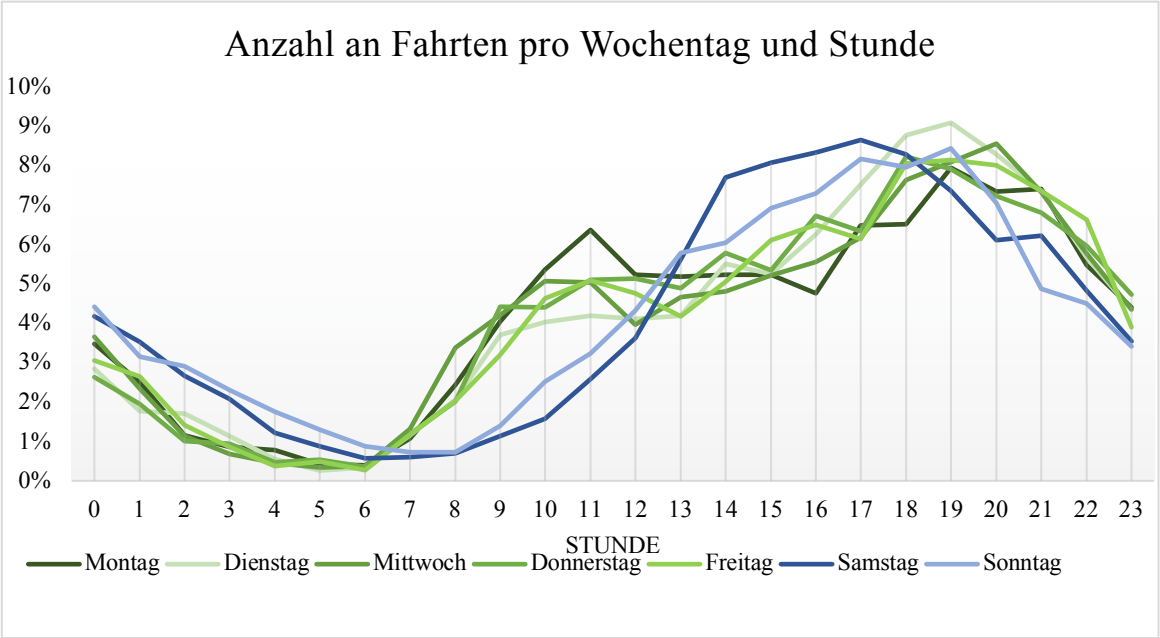

Abbildung 3: Anzahl an Fahrten pro Wochentag und Stunde

In Abbildung 4 sind die Anzahl der Fahrten pro Stunde in % aller Fahrten an Werktagen bzw. Wochenenden noch einmal zusammengefasst zu erkennen. Die meisten Fahrten starten mit circa 8% aller Rollermieten an Werktagen zwischen 19:00 und 19:59 Uhr, am Wochenende starten die meisten Fahren mit knapp 8% zwischen 17:00 und 17:59Uhr. An Werktagen finden die geringsten Fahrten zwischen 6:00 und 6:59Uhr und an Wochenenden zwischen 7:00 und 7:59Uhr statt, beide Werte machen jeweils weniger als 1% aller Rollerreservierungen im Erhebungszeitraum aus.

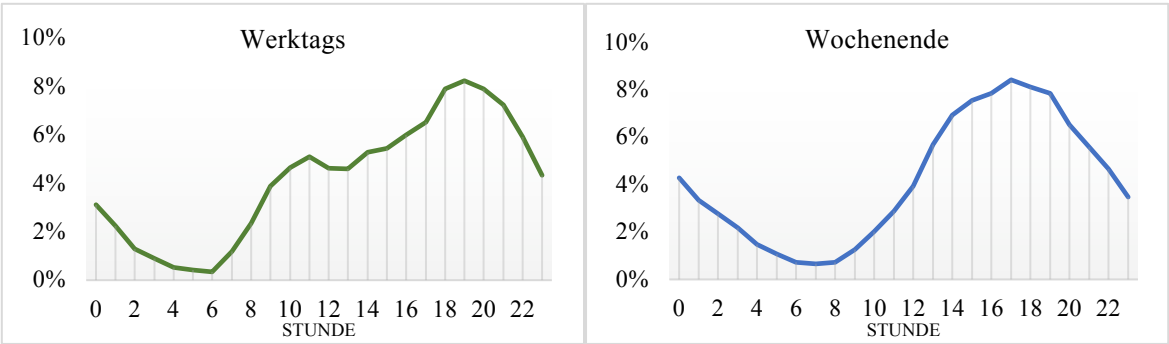

Abbildung 4: Anzahl an Fahrten – Werktags und Wochenende

Die Standzeiten der Roller nach der Stunde des Startzeitpunktes zeichnen sich dementsprechend spiegelbildlich zur Rollernachfrage ab, wie in Abbildung 5 dargestellt. Ein Roller, welcher als Startzeitpunkt der Standzeit die Stunde 0 hat, wurde also zwischen 00:00 und 00:59 abgestellt und zieht an einem Werktag im Durchschnitt 6,3h, am Wochenende nur 4,6h Standzeit nach sich. Das Minimum an durchschnittlicher Standzeit ist an Werktagen

zwischen 17:00 und 17:59Uhr mit 1,5h und am Wochenende im Zeitfenster zwischen 16:00 und 16:59Uhr mit 1,3h.

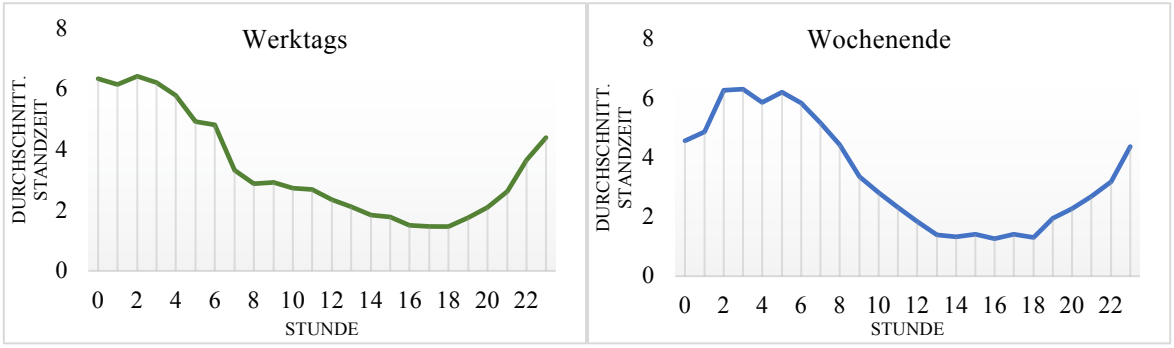

Abbildung 5: Verlauf der durchschnittlichen Standzeit - Werktags und Wochenende

3.3.2 Räumliche Verteilung der Standzeiten in den LOR

Zusätzlich zum zeitlichen Verlauf soll in diesem Abschnitt die räumliche Verteilung der Standzeiten auf LOR Ebene erläutert werden. Mit Hilfe der Open-Source Software QGIS werden die mittleren Standzeiten eines jeden LOR in Berlin zu bestimmten Zeitpunkten abgebildet. Abbildung 6 zeigt die mittlere Standzeit pro LOR an Werktagen bzw. Wochenenden. Je heller eine LOR gefärbt ist, desto höher ist die mittlere Standzeit, dh. in diesen Gebieten in Berlin stehen die Roller im Durchschnitt am längsten bevor sie wieder angemietet werden. Die dunklen LOR zeigen die beliebten Gebiete auf, zu denen vor allem die Bezirke Prenzlauer Berg, Kreuzberg, Mitte, Friedrichshain und Neukölln zählen. Zusätzlich zur farblichen Markierung geben die einzelnen Zahlen in jedem LOR die tatsächliche Anzahl an Datensträngen an, die sich zur mittleren Standzeit zusammensetzen. So hat ein LOR, das zwar dunkel gefärbt ist, dessen Berechnung der mittleren Standzeit jedoch lediglich auf wenigen Erhebungen beruht, folglich wenig Aussagekraft. Die Einteilung in sechs Klassen erfolgt anhand des Jenks-Caspall-Algorithmus, welcher die Unterschiede innerhalb einer Klasse minimiert bei gleichzeitiger Maximierung der Unterschiede zwischen den Klassen.

In Abbildung 6 sieht man deutlich, dass an den Wochenenden die mittlere Standzeit vor allem in den Wohngebieten, die in Abbildung 2 dargestellt sind, niedrig ist. An Werktagen verzeichnet zusätzlich das Gebiet im Zentrum und um den Bahnhof Zoologischer Garten im Westen eine niedrige mittlere Standzeit zwischen 1,12 und 2,20h auf.

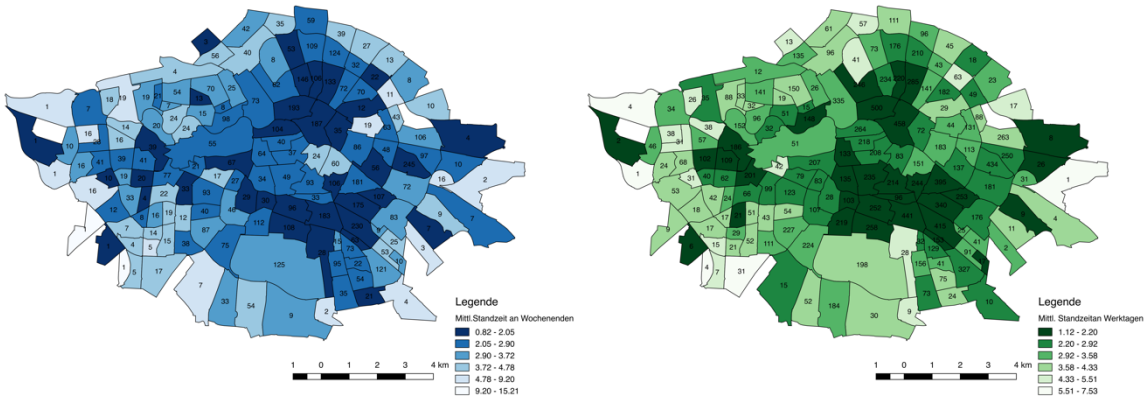

Abbildung 6: Mittlere Standzeit der LOR an Wochenenden und Werktagen

Abbildung 7 vergleicht die räumliche Verteilung der mittleren Standzeit zum Zeitfenster 9 und 17 Uhr an Werktagen. Die Grafik auf der linken Seite, welche das Zeitfenster 9 abbildet, zeigt auf, dass die kürzesten mittleren Standzeiten an den Rändern des Geschäftsgebietes zu verzeichnen sind. Diese LOR sind erneut diejenigen, in denen viele der registrierten emmy-Kunden wohnen. Im Zeitfenster 17 dagegen weisen zusätzlich die LOR im Zentrum kurze Standzeiten zwischen 0,01 und 1,07h auf.

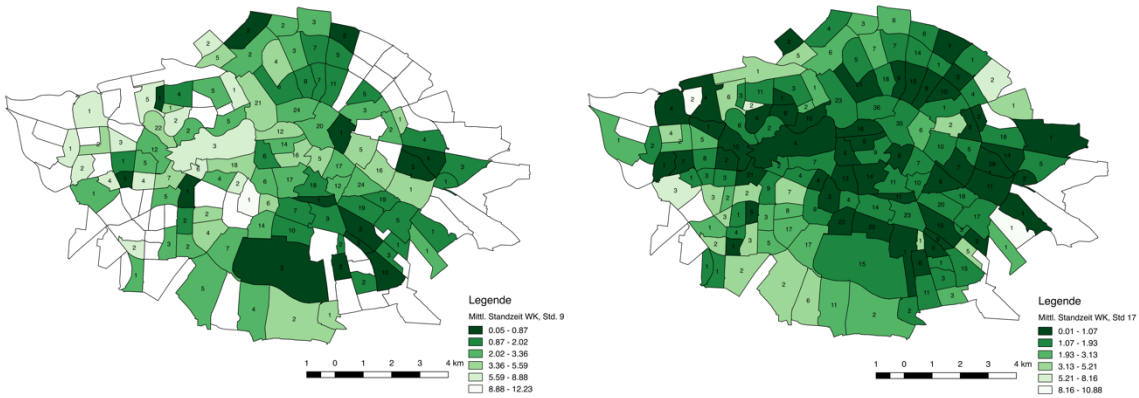

Abbildung 7: Mittlere Standzeit an Werktagen im Zeitfenster 9 und 17Uhr

3.3.3 Allgemeine deskriptive Statistik

Beim vorliegenden Datensatz handelt sich um die Grundgesamtheit aller erhobenen Standzeiten innerhalb von sechs Wochen. Die Standzeiten in Sekunden bilden die Grundlage des Datensatzes und sollen anhand externen Faktoren erklärt werden. Die einzelnen erklärenden Variablen sind in Tabelle 1 aufgelistet. Sie lassen sich in vier Klassen teilen: Meteorologische, sozioökonomische und zeitliche Faktoren und Roller-/Flotteneigenschaften. Für die metrische bzw. ordinale Variable *Niederschlag* und *Startzeitpunkt* wird im weiteren Verlauf jeweils eine binäre Dummy-Variablen einführt. Beträgt die Dummy-Variable *dummyNiederschlag* den Wert 1, dann hat es innerhalb der Starzeitpunktstunde, also zum Zeitpunkt der Abmietung eines Rollers, geregnet. Der Wert 0 steht dementsprechend für keinen Niederschlag. Außerdem werden Dummy-Variablen für jedes einzelne Stundenfenster des Startzeitpunktes eingeführt, welche mit beispielsweise *Stunde_13* bezeichnet werden. Als Startzeitstunde gilt der Zeitpunkt, an dem der Roller abgestellt wurde und somit wieder für alle Kunden in der App sichtbar ist. Ein Roller, der beispielsweise um 13:15Uhr abgestellt wurde, wird dementsprechend der Startzeitstunde 13 zugeordnet. Die Variable *Fahrzeugtyp* gibt an, um welches Rollermodell es sich handelt. Insgesamt bietet emmy vier verschiedene Modelle an, die sowohl von unterschiedlichen Herstellern produziert wurden, als auch unterschiedlich zu bedienen sind. Es wird daher vermutet, dass die Kunden ein Rollermodell einem anderen vorziehen könnten. Die sozioökonomischen Daten wurden anhand des Standortes des Rollers mit den Eigenschaften der LOR-Flächen verknüpft. Die Variable *lorEinwohner* berücksichtigt die Bevölkerung im Alter von 18-85 Jahren, da diese aus unternehmenseigener Angabe die Zielgruppe von emmy abbilden. Die Daten zur Anzahl der Haltestellen wurden anhand des Datensatzes „Rohdaten: Berlin an deiner Linie“ der Berliner Morgenpost mit emmys Nutzerdaten verknüpft (Berliner Morgenpost, 2017).

Tabelle 2 zeigt die wichtigsten Lage- und Streuparameter der Variablen auf. Die sozioökonomischen Faktoren der LOR-Gebiete konnten anhand fehlender Daten nicht dem kompletten Datensatz zugeordnet werden, weswegen die Anzahl der einzelnen Variablen stark variierende Werte aufzeigt. So sind zwar 23.776 Datenstränge mit Angaben zur Standzeit, Wetter und den Roller bzw. Rollereigenschaften vorhanden, die Anzahl der Daten zur Gewerbenutzung in einem LOR beispielsweise beträgt jedoch nur 13.992. Die mittlere Standzeit beträgt 9466,83 Sekunden (ca. 2,6h), wobei die Variable eine große Standardabweichung von 12.310,46 um den Mittelwert vorweist. Die Temperatur hat eine Spannweite von 4,9 bis hin zu 23,10 Grad und leidglich 6% der ermittelten Daten zeichnen

Niederschlag auf. Im Durchschnitt waren mit nur kleineren Abweichungen ca. 48% der gesamten Flotte täglich aktiv. 70% des Datensatzes sind Standzeiten, die an Werktagen ermittelt wurden.

Variablen	Definition	
T_t	Temperatur in °C zur Startzeitstunde t	*Meterologi.Faktore*
NS_t	Dummy-Variable zum Niederschlag zur Startzeitstunde t $NS_t = 0: Kein\ Niederschlag;$ $NS_t = 1: Niederschlag$	
WS_t	Windstärke in km/h zur Starzeitstunke t	
FT	Fahrzeugtyp des Rollers: $FT = 11: Nuvi;$ $FT = 13: Nuva;$ $FT = 15: Muvi;$ $FT = 18: Schwalbe$	*Roller-/ Flotteneigenschaften*
AS	Akkustand in % des abgestellten Rollers	
AR	Anteil der aktiven Roller der Gesamtflotte in Berlin in %	
WN_{lor}	Anteil der Wohnnutzung im LOR des abgestellten Rollers	*Sozioökonomische Faktoren*
GN_{lor}	Anteil der gewerblichen Flächennutzung im LOR des abgestellten Rollers	
VF_{lor}	Anteil der Verkehrsfläche (Straßen, Schienen etc.) im LOR des abgestellten Rollers	
EW_{lor}	Anzahl der Einwohner im Alter von 18-85Jahren im LOR / Fläche des LOR in km^2	
K_{lor}	Anzahl der registrierten emmy Kunden im LOR / Fläche des LOR in km^2	
$\ddot{O}PNV_{lor}$	Anzahl der ÖPNV-Haltestellen im LOR / Fläche des LOR in km^2	
S_t	Dummy-Variable zur jeweiligen Startzeitstunde zur Basis 0 Es gilt: $S = 0:$ Roller wurde nicht in dieser Startzeitstunde abgestellt $S = 1:$ Roller wurde in dieser Startzeitstunde abgestellt	*Zeitliche Faktoren*

Tabelle 1: Definition der erklärenden Variablen

Deskriptive Statistik

	N	Spannweite	Min.	Max.	Mittelwert	Standardfehler	Standardabw.	Varianz
Standzeit	23776	55046	22	55068	9466,83	79,837	12310,457	151547348,60
logStandzeit	23776	7,83	3,09	10,92	8,1826	,01077	1,66115	2,759
Startzeitpunkt	23776	86388	1	86389	54647,05	142,630	21992,72	483679914,80
Werktags	23776	1	0	1	,70	,003	,459	,211
Temperatur	23776	18,20	4,90	23,10	14,3893	,01971	3,03914	9,236
Niederschlag	23776	5,2	,0	5,2	,047	,0018	,2777	,077
dummyNiederschlag	23776	1	0	1	,06	,002	,243	,059
Wind	23776	9,00	,30	9,30	3,1391	,00975	1,50373	2,261
Fahrzeugtyp	23776							
Akkustand	23776	255	0	255	60,55	,149	22,932	525,860
aktiveRoller	22875	,22	,35	,57	,4832	,00043	,06482	,004
lorWohnnutzung	22886	,67	,00	,67	,2979	,00100	,15137	,023
lorGewerbenutzung	13992	,63	,00	,63	,0928	,00077	,09139	,008
lorVerkehrsfläche	17843	,25	,00	,25	,0480	,00040	,05389	,003
lorEinwohnerFlache	23602	29572,55	,00	29572,55	12808,0547	43,44564	6674,52876	44549334,170
lorKundenFlaeche	23602	576,81	,00	576,81	203,4620	,94882	145,76657	21247,893
lorÖPNV	22940	104,09	,00	104,09	19,8891	,10103	15,30221	234,158

Tabelle 2: Deskriptive Statistik

Im weiteren Verlauf der Analyse wird die Variable *Standzeit* statistisch durch Logarithmieren standardisiert, um sie anschließend in einer linearen Regression abbilden zu können. Durch das Logarithmieren wird eine spätere unrealistische Prognose von negativen Standzeitenwerten durch die Regressionsmodelle verhindert. Abbildung 8 zeigt die Verteilung von *logStandzeit* in einem Histogramm, wobei deutlich zu erkennen ist, dass sie keiner Normalverteilung unterliegt, was auch durch Abbildung 9 verdeutlicht wird. Der Boxplot in Abbildung 8 zeigt auf, dass Werktage und Wochenende ähnliche Streuung der Variable *logStandzeit* vorweisen. Die durchschnittliche Standzeit schwankt nur gering mit einem logarithmierten Wert von 8,23 (ca. 2,67h) an Werktagen und 8,08 (ca. 2,55h) an Wochenenden.

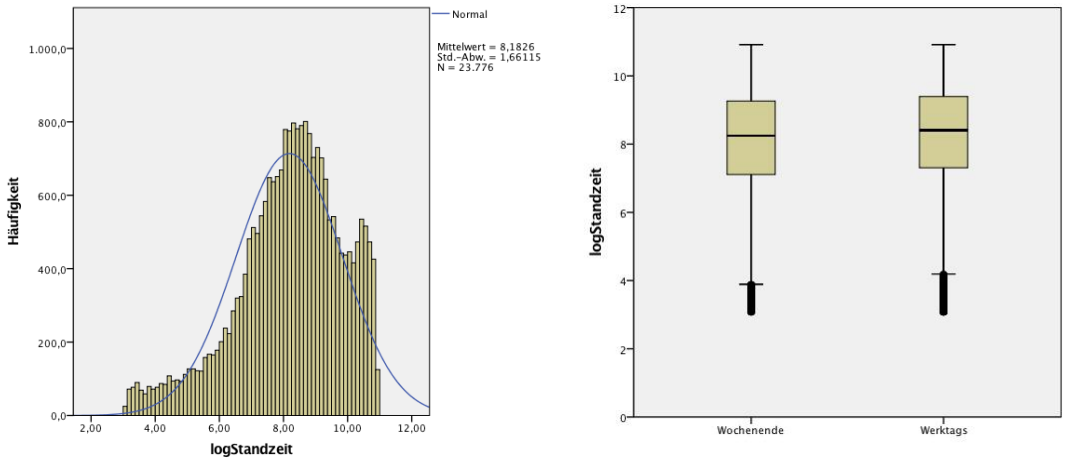

Abbildung 8: Histogramm der logStandzeit und Boxplot der logStandzeit an Werktagen und Wochenenden

Starke Abweichungen von der Geraden mit 45° Steigung im Quantil-Quantil Plot sind am linken und rechten Rand in Abbildung 9 zuerkennen. Im Q-Q Plot werden die empirischen Quantile gegen die theoretischen Quantile der Verteilung zur Normalverteilungsprüfung abgebildet. Es lässt sich jedoch eine Verbesserung im zweiten Q-Q Plot der Variable *logStandzeit* gegenüber dem Q-Q Plot der untransformierten Variable *Standzeit* feststellen.

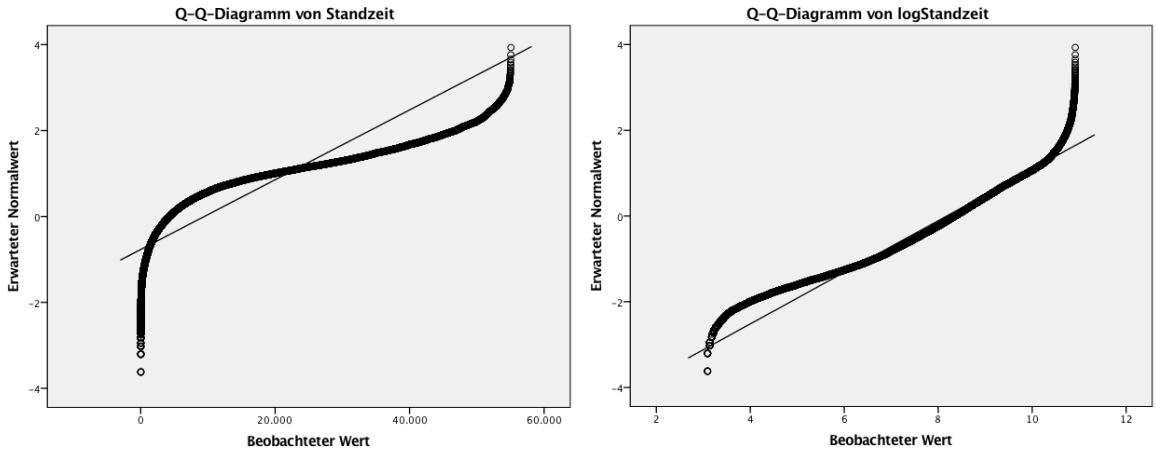

Abbildung 9: q-q Plot von Standzeit und logStandzeit

4. Methoden

In diesem Kapitel wird die statistische Methode der Standzeitenanalyse vorgestellt, die als Werkzeug zur Durchführung und späteren Interpretation der Forschungsergebnisse dient. Um das Modell zur Standzeitenanalyse umfassend zu erforschen, wurden zwei unterschiedliche Verfahren zur Aufstellung von Regressionsmodellen angewendet. Im Folgenden wird zunächst die Vorgehensweise und Theorie der linearen Regression erläutert, im zweiten Abschnitt folgt die statistische Methode der Cox-Regression.

4.1 Multiple lineare Regression

Ziel einer multiplen linearen Regression ist es, eine abhängige Zielvariable anhand mehreren unabhängigen Variablen zu erklären. Der Einfluss der unabhängigen Variablen auf die Zielvariable sollte linear sein, Nichtlinearität kann durch passende Transformation des nichtlinearen Modells in ein lineares Modell überführt werden. In einem linearen Modell setzt sich die zu erklärende Variable aus einer Linearkombination der Regressionskoeffizienten zusammen. Anhand der Methode der kleinsten Quadrate, die im zweiten Abschnitt erläutert wird, lassen sich die Regressionskoeffizienten berechnen. Im folgenden Abschnitt wird zunächst ein allgemeiner Überblick über die Hauptfaktoren einer linearen Regression gegeben.

Die vorliegende Arbeit folgt in der Erklärung der linearen Regression dem Buch „Regression: Modelle, Methoden und Anwendungen“ von Fahrmeir, L., Kneib, T. und Lang, S. aus dem Jahr 2009.

4.1.1 Aufstellen von linearen Regressionsmodellen

Der Einfluss der Variable X auf die primäre, metrische Variable Y lässt sich anhand der Funktion $f(x_{1i}, \dots, x_{ki})$ wie folgt darstellen:

$$y_i = f(x_{1i}, \dots, x_{ki}) + \varepsilon_i, \qquad i = 1, \dots, n. \tag{4.1}$$

Da in der Praxis der funktionale Zusammenhang zwischen y_i und den zu erklärenden Variablen nicht zu einer absoluten Erklärung führt, nimmt man einen weiteren Störterm ε_i an.

Die unbekannte Funktion $f(x_{1i}, \dots, x_{ki})$ setzt sich aus einer Linearkombination der Regressoren zusammen, d.h.

$$f(x_{1i}, \dots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}. \tag{4.2}$$

Die Parameter $\beta_1, \dots, \beta_k$ sind die unbekannten Regressionskoeffizienten der exogenen Variablen und der Parameter β_0 dient als Konstante. Beide Parameter werden anhand der Methode der kleinsten Quadrate oder anderen Verfahren geschätzt.
Für ein multiples lineares Regressionsmodell werden folgende Annahmen über die Störgrößen getroffen. Die Störgrößen haben den Erwartungswert Null, d.h. es gilt

$$E(\varepsilon) = 0. \tag{4.3}$$

Außerdem wird angenommen, dass die Störgrößen unkorreliert und die Varianz der Störungen für alle Beobachtungen i konstant ist, d.h. die Kovarianzmatrix entspricht

$$Cov(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I. \tag{4.4}$$

Ist die zweite Annahme über die Störgrößen (4.4) nicht erfüllt, spricht man von heteroskedastischen Fehlern. Zusätzlich muss bei einem linearen Regressionsmodell die Annahme gelten, dass die Designmatrix X, welche alle Werte der unabhängigen Variable umfasst, vollen Spaltenrang besitzt, d.h.

$$rg(X) = k + 1 = p. \tag{4.5}$$

Die Spalten von X sind also linear unabhängig, was die notwendige Bedingung stellt, dass die Anzahl n an Beobachtungen größer oder gleich der Anzahl der Regressionskoeffizienten p ist. Zuletzt wird angenommen, dass die Störgrößen normalverteilt ist.

4.1.2 Methode der Kleinsten Quadrate

Die Regressionsparameter β_k sind unbekannt und werden anhand der Methode der Kleinsten Quadrate nach dem Mathematiker Adrian Marie Legendre aus dem Jahr 1805 geschätzt. Diese Methode beruht auf der Minimierung der Summe der quadratischen Abweichung zur Schätzung von β_k. Dabei werden also die Residuen, welches den Abstand zwischen der Zielvariable y_i und den Regresswerten $\hat{y}_i$ bezeichnet, so klein wie möglich gehalten. Es gilt

$$KQ(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon. \tag{4.6}$$

bezüglich $\beta \in \mathbb{R}^p$ mit $p := k + 1$.
Durch das Nullsetzen der Gleichung (4.6) wird das Minimierungsproblem gelöst. Der Regressionsschätzer $\hat{\beta}$, den man durch diese Methode erhält, setzt sich aus der sogenannten Normalgleichung zusammen, d.h.

$$\hat{\beta} = (X'X)^{-1}X'y. \quad (4.7)$$

$X'X$ ist positiv definit und damit invertierbar, was zu einer eindeutigen Lösung der Normalgleichung führt.

4.1.3 Bewertung von linearen Regressionsmodellen

Anhand einer Einflussanalyse lassen sich einzelne Beobachtungen, die einen großen Einfluss auf $\hat{\beta}$ bzw. $\hat{y}$ haben, aus dem Datensatz herausfiltern. Dieses kann anhand der Berechnung des euklidischen Abstandes bzw. der Cooks-Distanz zwischen den beiden Schätzungen $\hat{y}$ und $\hat{y}_{(i)}$ erfolgen. $\hat{y}_{(i)}$ bezeichnet eine Schätzung, die auf allen Beobachtungen bis auf der *i*-ten beruht. Der Abstand der Schätzungen ist mit der geschätzten Varianz $\hat{\sigma}^2$ gewichtet, d.h.

$$D_i = \frac{(\hat{y}_{(i)}-\hat{y})'(\hat{y}_{(i)}-\hat{y})}{p\cdot\hat{\sigma}^2}. \quad (4.8)$$

Laut Fahrmeir et al. gilt, dass Beobachtungen mit $D_i > 0{,}5$ als auffällig gelten, Beobachtungen mit $D_i > 1$ sollten näher betrachtet werden.

Um die Modellannahmen aus Abschnitt 4.1.1 eines linearen Regressionsmodells zu überprüfen, werden verschiedene Vorgehensweisen wie beispielsweise dem Test auf Multikollinearität durchgeführt. Ein Multikollinearitätsproblem liegt vor, wenn zwei oder mehrere erklärende Variablen stark miteinander korrelieren. Die Aussagen über die Schätzung der Regressionskoeffizienten wird somit ungenau und das Modell kann nicht mehr eindeutig interpretiert werden. Ob und in welchem Ausmaß eine Korrelation zwischen den erklärenden Variablen vorliegt, lässt sich anhand des Varianzinflationsfaktors VIF_j überprüfen, d.h.

$$VIF_j = \frac{1}{1-R_j^2}. \quad (4.9)$$

mit R_j als Bestimmtheitsmaß der Regression von x_j auf alle übrigen Einflussgrößen. Der Varianzinflationsfaktor gibt an, um welchen Faktor die Varianz von $\widehat{\beta_j}$ durch die lineare Abhängigkeit mit anderen erklärenden Variablen korreliert. Sowohl das Bestimmtsheitsmaß R_j^2, welches im Folgenden näher erläutert wird, als auch der VIF_j sind bei stärkerer Abhängigkeit größer, bei einem $VIF_j > 10$ liegt ein erstaftes Kollinearitätsproblem vor.

Als Maß für die Güte der Anpassung eines Regressionsmodells an die Daten kann das Bestimmtsheitsmaß betrachtet werden. Der Bestimmtheitskoeffizient R^2 gibt Auskunft über die Aussagekraft eines Modells, d.h. den Anteil der Varianz, der durch das Regressionsmodell erklärt werden kann. R^2 ist auf das Intervall [0,1] normiert. Definiert ist das Bestimmtheitsmaß durch

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad (4.10)$$

mit $\bar{y}$ als Mittelwert der Zielgröße y. Je näher R^2 bei 1 liegt, desto besser ist das Modell und die Varianz der Zielgröße kann durch die Regressoren erklärt werden. Umgekehrt ist die Anpassung der Daten an das Modell gering, wenn R^2 nahe 0 liegt. Im Extremfall ist $R^2 = 1$, d.h. alle Residuen sind gleich Null und die Anpassung der Daten an das Modell ist daher perfekt. Da das Bestimmtheitsmaß bei einer zusätzlichen Kovariablen automatisch größer wird, verwendet man das sogenannte korrigierte Bestimmtheitsmaß, welches definiert ist durch

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2). \qquad (4.11)$$

Zusätzlich zu der Ermittlung des Bestimmtheitmaßes und dem Test auf Multikollinearität ist es sinnvoll, eine Residuenanalyse durchzuführen. Anhand der Residuenanalyse lassen sich weitere Verletzungen der Annahmen über den Störterm ε wie beispielsweise Homoskedistizität aufdecken. Eine graphische Analyse der Residuenplots ist dabei die aussagekräftigste Methode. Bei einer solchen Analyse werden die vorhergesagten Werte $\hat{y}$ gegen die geschätzten Residuen $\hat{\varepsilon}$ geplottet. Die Punkte sollten im Diagramm unsystematisch streuen, um die Annahme einer Heteroskedistizität verwerfen zu können.

4.2 Cox-Regression

Die zweite Methodik, die im fünften Kapitel angewendet wird, ist die Cox-Regression, welches eine spezielle Form der Ereignisanalyse ist. Sie ist eng verwandt mit multiplen bzw. logistischen Regressionsmodellen und untersucht die Länge eines Zeitintervalls bis zum Eintreten eines bestimmten Ereignisses. Die sogenannte Überlebenswahrscheinlichkeit bzw. Hazard-/Ausfallrate wird unter Einfluss verschiedener metrischer oder kategorialen Kovariablen untersucht.

Die Cox-Regression ist die populärste Form eines Überlebenszeitenmodells bzw. Hazard Modell, wobei das Modell sehr häufig in der Erforschung demographischer Daten wie beispielsweise der Untersuchung des Therapieeffektes von Medikamenten verwendet wird. Es handelt sich um ein semiparametrisches Verfahren, das 1972 von David Cox entwickelt wurde. Im Gegensatz zur linearen Regression aus Abschnitt 4.1 ist die Verwendung der Partial-Likelihood-Methode , welche in Abschnitt 4.2.2 erläutert wird, zur Schätzung der Regressionsparameter üblich.

Die theoretische Darstellung der Cox-Regression basiert auf der Veröffentlichung „Multivariate Statistische Verfahren“ von Fahrmeir, L., Hamerle, A. & Tutz, G. (1996).

4.2.1 Allgemeine Theorie zur Cox-Regression

Das Cox-Modell wird bei der Untersuchung des gleichzeitigen Effektes von mehreren Zufallsvariablen auf die Zielvariable eingesetzt. Viele Modelle weisen unvollständige Datensätze auf, in denen die Verweildauer bzw. die Überlebenszeit nach Beendigung der Beobachtung möglicherweise noch weiter anhält. In solchen Fällen werden verschiedene Varianten der Zensierung angewendet. Da in dieser Arbeit der relevante Datensatz vollständig ist und daher nicht zensiert werden muss, wird auf die Erläuterung von Zensierungsverfahren verzichtet.

Die Zielvariable wie beispielsweise die Überlebenszeit eines Individuums muss bei einer Cox-Regression keiner bestimmten Verteilung unterliegen. Vorausgesetzt wird aber, dass die Effekte der Variablen auf die Zielvariable bzw. die Hazard-Rate über die Zeit hinweg konstant sind. Die Hazardrate oder Ausfallrate $h_i(t)$ ist definiert als bedingte Wahrscheinlichkeit relativ zur Breite des Zeitintervalls Δt., dass das Ereignis im Zeitintervall $[t, t + \Delta t]$ auftritt, unter der Voraussetzung, dass das Ereignis bis zur Zeit t noch nicht stattgefunden hat, d.h.

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_i < t + \Delta t | T_i \geq t)}{\Delta t}. \qquad (4.12)$$

Die Hazardrate $h_i(t)$ ist unter Einfluss von Kovariaten β_k modelliert als Produkt einer unspezifizierten Baseline-Funktion $h_0(t)$ und einem weiteren Term, der den Einfluss eines Kovariatvektors $X_{ik}(t)$ angibt. Die unspezifizierte Baseline-Funktion $h_0(t)$ muss in jedem Fall positiv sein und hat ähnliche Bedeutung wie der Achsenabschnitt einer linearen Regression. Sind alle Einflussvariablen gleich Null, gibt $h_0(t)$ die Hazardrate an. Allgemeint ist $h_i(t)$ wie folgt modelliert:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 x_{1i} + \ldots + \beta_k x_{ki}). \quad (4.13)$$

Die Regressoren β_k beeinflussen die Hazardrate also proportional.
Für jedes Paar von Individuen gilt zu jedem Zeitpunkt t, dass

$$\frac{h_i(t)}{h_j(t)} = c \text{ mit } i = 1, \ldots, n\ , \quad (4.14)$$

wobei c zwar von den erklärenden Variablen abhängen kann, nicht jedoch von der Zeit t. Allgemein gilt, dass der erste Term der Gleichung (4.13) $h_0(t)$ nur von der Zeit abhängig ist, der zweite Term hingegen von dem Kovariatvektor $X_{ki}(t)$.

4.2.2 Partial-Likelihood Methode und Umgang mit Ties

Die Regressionsparameter β_k werden anders als bei der linearen Regression in Abschnitt 4.1 durch die Methode des Partial-Likelihood geschätzt. Die Partial-Likelihood PL wird wie eine gewöhnliche Likelihood-Funktion behandelt, mit dem Unterschied, dass die Likelihood faktorisiert wird. Es ergibt sich folgende Formel für PL, die in Abhängigkeit von β maximiert wird:

$$PL(\beta, x_1, \ldots, x_N) = \prod_{n=1}^{k} \frac{\exp(x_n'\beta)}{\sum_{l \in \mathbb{R}_{t_n}} \exp(x_l'\beta)}. \quad (4.15)$$

Die Risikomenge $\mathbb{R}_{t_n}$ beinhaltet die Individuen, bei denen das Ereignis vor dem Zeitpunkt t noch nicht eingetreten ist und die durch die Zensierung noch nicht exkludiert wurden.
Trifft das Ereignis für mindestens zwei Individuen gleichzeitig ein, spricht man von Ties. Im Falle von Ties in einer Beobachtung können die Methode der Partial-Likelihood nur unter weiteren Anpassungen durchgeführt werden. Zu den verbreiteten Methoden zählen Breslow (1974) und Efron (1977). Nach der Methode von Breslow(1974) wird die Gleichung (4.15) durch die Approximation korrigiert, d.h.

$$PL(\beta, x_1, \ldots, x_N) = \prod_{n=1}^{k} \frac{\exp(s_n'\beta)}{\left[\sum_{l \in \mathbb{R}_{t_n}} \exp(x_l'\beta)\right]^{d_n}}. \quad (4.16)$$

Die Anzahl der Ties zum Zeitpunkt t_n wird somit durch d_n berücksichtigt. s_n ist die Summe der Kovariablenvektoren aller d_n.

4.2.3 Tests der Regressionsparameter

Im letzten Schritt der Modellierung einer Cox-Regression müssen die Regressionsparameter getestet und nicht signifikante erklärende Variablen aus dem Modell ausgeschlossen werden. Die Wald-Statistik gibt dabei die Signifikanz des geschätzten Koeffizienten β an, welches die gleiche Bedeutung wie der t-Test in einem multiplen Regressionsmodell hat. Mit Hilfe der Methode einer Schrittweisen-Selektion werden Regressoren anhand ihres Signifikanzwertes der Wald-Statistik zum Modell hinzugefügt oder ausgeschlossen.
Als Maß für die Güte der Anpassung eines Modells an die Daten wird der negative doppelte Wert des Logarithmus der Likelihood-Funktion -2LL verwendet, welcher auch beim Verfahren der logistischen Regression Anwendung findet. Durch Hinzufügen oder Eliminierung von Kovariablen ändert sich der Wert von -2LL und nimmt Werte zwischen 0 und $+\infty$ an. Je weiter -2LL von 0 entfernt ist, desto schlechter ist das Modell an die Daten angepasst.

5. Regressionsergebnisse und Interpretation

In der multiplen linearen Regression ist die zu erklärende Variable Y die logarithmierte Standzeit. Diese Transformation der Standzeit in Sekunden ist sinnvoll, um letztendlich ein Modell aufzustellen, dass nur positive Werte für die Standzeit liefert.

Das zweite Modell, die Cox-Regression, wiederum verwendet die Standzeit in Sekunden in untransformierter Form als Zeitvariable, um möglichst wenige Ties im Datensatz aufzuweisen. Die erklärenden Variablen beziehen sich auf die einzelnen Standzeiten und wurden teilweise aus anderen Datenquellen stammend mit dem Datensatz der Standzeiten verknüpft. Einen Überblick aller erklärenden Variablen und deren Definition ist in Tabelle 1 aus Abschnitt 3.3.3 zu finden.

Im weiteren Verlauf der Analyse werden verschiedene Methoden angewandt, um das Modell zu finden, welches anhand von Kovariablen die Standzeit eines Rollers bzw. das Risiko einer langen Standzeit am besten erklärt. In Abschnitt 5.1 werden zunächst die Ergebnisse des linearen Regressionsmodells vorgestellt, gefolgt von den Ergebnissen der Cox-Regression in Abschnitt 5.2.

5.1 Interpretation der Ergebnisse des linearen Regressionsmodells

Zunächst wird der Zusammenhang zwischen den einzelnen erklärenden Variablen und der Variable *logStandzeit* erläutert und anhand einer Korrelationsmatrix und Streudiagrammen dargestellt. Dann erfolgten die Aufstellung des Regressionsmodells und die Selektion der erklärenden Variablen. Im letzten Abschnitt wird die durchgeführte Residuenanalyse erläutert.

5.1.1 Zusammenhangsanalyse und Modellaufstellung

Anhand der Spearsons-Rangkorrelation wird der Grad des Zusammenhanges zwischen den einzelnen numerischen Variablen berechnet und die unabhängigen Variablen auf Multikollinearität überprüft. Indikator für Multikollinearität ist eine sehr hohe positive oder negative Korrelation zweier Regressoren, d.h. r_{SP}weist Werte nahe 1 bzw. -1 auf. Eine hohe Korrelation zwischen einem Regressor und der zu erklärenden Variablen Y ist dagegen von Vorteil. Weist r_{SP} Werte nahe oder gleich Null auf, besteht ein schwacher bzw. garkein Zusammenhang der Variablen. Es zeigt sich sowohl eine starke Korrelation r_{SP}=0,849 der Variablen EW_{lor} und WN_{lor} als auch zwischen EW_{lor} und K_{lor} mit r_{SP}=0,854. Zusätzlich

zur Korrelationsmatrix wird der Varianzinflationsfaktor VIF für die Variablen aus der Korrelationsmatrix berechnet (siehe Anhang A). $VIF_{EW_{lor}}$ beträgt 13,214 und zählt damit laut Daumen-Regel nach Fahrmeir (2009) als Kollinearitätsproblem. Die Beibehaltung von K_{lor} begründet sich auch mit der stärkeren Anpassung an emmys Kundenbasis und wird daher über die absolute Einwohnerdichte pro LOR bevorzugt im Modell beibehalten. Alle anderen Werte des VIF sind kleiner 10 und geben keinen Grund zur näheren Betrachtung. Tabelle 3 zeigt außerdem auf, welche Variablen den stärksten Einfluss auf die zu erklärende Variable *logStandzeit* haben. Am stärksten korreliert die Startzeitstunde S_t mit der Standzeit (r_{SP}= -0,176), gefolgt von der Temperatur in °C, welche mit einem Wert von r_{SP}= -0,154 mit der logarithmierten Standzeit korreliert. Die Kundendichte pro LOR hat eine negative Korrelation mit *logStandzeit* von r_{SP}= -0,106.

		logStandzeit	Startzeitpunkt_Stunde	Temperatur	Wind	Akkustand	aktiveRoller	lorWohn-nutzung	lorGewerbe-nutzung	lorVerkehrs-fläche	lorEinwohner Flache	lorKunden Flaeche	lorÖPNV
$logST$	r_{SP}	1,00	-,176	-,154	-,002	,030	,028	-,023	,023	,094	-,058	-,106	-,018
	Sig.	.	,000	,000	,804	,000	,000	,001	,006	,000	,000	,000	,006
S_t	r_{SP}	**-,176**	1,00	-,148	-,143	-,092	,033	,067	-,020	-,014	,081	,078	-,001
	Sig.	,000	.	,000	,000	,000	,000	,000	,020	,061	,000	,000	,912
T_t	r_{SP}	**-,154**	-,148	1,00	,269	,046	-,063	-,107	,012	,026	-,129	-,119	,024
	Sig.	,000	,000	.	,000	,000	,000	,000	,169	,001	,000	,000	,000
WS_t	r_{SP}	-,002	-,143	,269	1,00	,063	,031	-,057	,000	,001	-,060	-,044	,024
	Sig.	,804	,000	,000	.	,000	,000	,000	,994	,871	,000	,000	,000
AS	r_{SP}	,030	-,092	,046	,063	1,00	,030	-,047	,016	,021	-,057	-,064	,014
	Sig.	,000	,000	,000	,000	.	,000	,000	,057	,004	,000	,000	,039
AR	r_{SP}	,028	,033	-,063	,031	,030	1,00	,012	,010	-,006	,019	,028	-,010
	Sig.	,000	,000	,000	,000	,000	.	,075	,248	,420	,003	,000	,140
WN_{lor}	r_{SP}	-,023	,067	-,107	-,057	-,047	,012	1,00	,040	-,348	,849	,672	-,178
	Sig.	,001	,000	,000	,000	,000	,075	.	,000	,000	,000	,000	,000
GN_{lor}	r_{SP}	,023	-,020	,012	,000	,016	,010	,040	1,00	,076	-,081	-,133	-,243
	Sig.	,006	,020	,169	,994	,057	,248	,000	.	,000	,000	,000	,000
VF_{lor}	r_{SP}	,094	-,014	,026	,001	,021	-,006	-,348	,076	1,00	-,385	-,359	-,007
	Sig.	,000	,061	,001	,871	,004	,420	,000	,000	.	,000	,000	,349
EW_{lor}	r_{SP}	-,058	,081	-,129	-,060	-,057	,019	**,849**	-,081	-,385	1,00	**,854**	-,120
	Sig.	,000	,000	,000	,000	,000	,003	,000	,000	,000	.	,000	,000
K_{lor}	r_{SP}	**-,106**	,078	-,119	-,044	-,064	,028	,672	-,133	-,359	,854	1,00	-,179
	Sig.	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	.	,000
$ÖPNV_{lor}$	r_{SP}	-,018	-,001	,024	,024	,014	-,010	-,178	-,243	-,007	-,120	-,179	1,00
	Sig.	,006	,912	,000	,000	,039	,140	,000	,000	,349	,000	,000	.

Tabelle 3: Korrelationsmatrix

Grafisch wird die Korrelation der Variable *Y* mit den erklärenden Variablen anhand von Streudiagrammen dargestellt. Die insgesamt niedrige Korrelation aller erklärenden Variablen mit der logarithmierten Standzeit, welche bereits in Tabelle 3 ersichtlich ist, wird in den Streudiagrammen nochmals deutlich. Lediglich das Streudiagramm von *logStandzeit* und *Temperatur* in Abbildung 10 weist ein zu erkennendes Muster auf, wodurch man einen negativen linearen Zusammenhang erkennen kann. Das Streudiagramm von *logStandzeit* und *Akkustand* in Abbildung 10 weist einen wesentlich geringeren linearen Zusammenhang auf, die eingezeichnete Regressionsgerade hat eine sehr flache Steigung. Aufgrund der hohen Datenmengen und der Überlappung von vielen Datenpunkten wurde die Variante des Sunflower-Plots zur Darstellung der Streudiagramme angewendet.

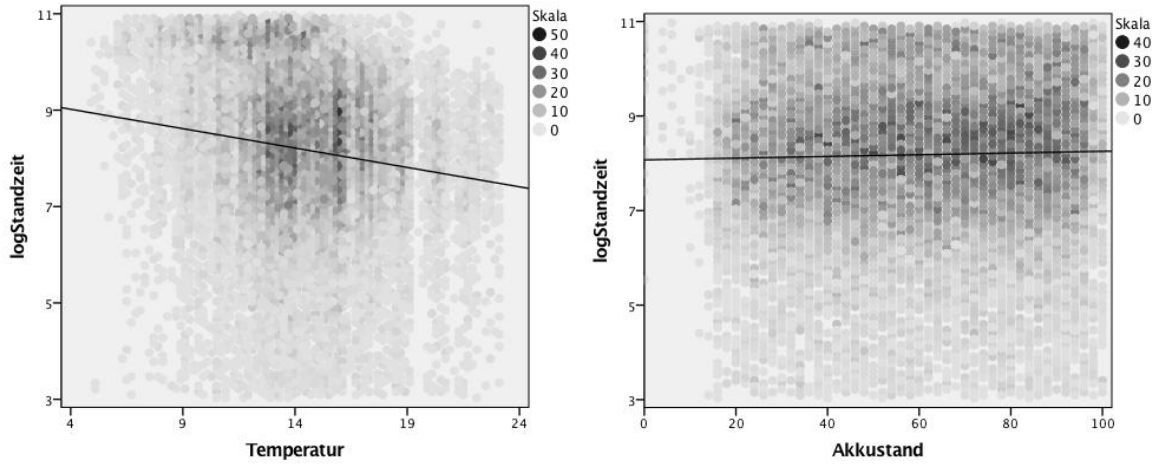

Abbildung 10: Streudiagramm zwischen logStandzeit und Temperatur bzw. Akkustand inkl. Regressionsgeraden

Trotz der geringen Korrelationen werden im nächsten Schritt nun Regressionsmodelle aufgestellt, um die tatsächlichen Regressionskoeffizienten zu schätzen und so eine Aussage über den Einfluss der Variablen X auf die Standzeit zu treffen. Aufgrund der großen Datenmengen und den unterschiedlichen zeitlichen als auch geographischen Verteilungen der Standzeit an Werktagen im Vergleich zu Wochenenden (siehe Abbildung 5 und 6), wird der Datensatz im weiteren Verlauf in zwei Gruppen aufgeteilt. Auf diese Weise kann eine genauere Aussage über die Ergebnisse der Regressionsmodelle getroffen werden.

Mit allen nicht ausgeschlossenen Variablen wird zunächst eine Regression durchgeführt und anschließend die Cook's Distanz aller Beobachtungen berechnet. Das Ergebnis zeigt keine auffälligen Werte $D_i > 0{,}5$, wodurch eine gesonderte Untersuchung laut Fahrmeir (2009) nicht notwendig ist. Da bereits Ausreißer anhand des Anomalieindex berechnet und entfernt wurden, dient die Berechnung der Cook's Distanz lediglich als weitere Überprüfung.

Das Modell, dass alle zur Verfügung stehenden Variablen beinhaltet, ist in der Theorie wie folgt definiert:

$$\begin{aligned} logST_{t,lor} = {} & \beta_0 + \beta_1 T_t + \beta_2 NS_t + \beta_3 WS_t \\ & + \beta_4 FT + \beta_5 AS + \beta_6 AR \\ & + \beta_7 WN_{lor} + \beta_8 GN_{lor} + \beta_9 VF_{lor} + \beta_{10} K_{lor} + \beta_{11} \text{ÖPNV}_{lor} \\ & + \beta_{12_{t=1}} S + \ \ldots \ + \beta_{25_{t=23}} S + \varepsilon_i . \end{aligned} \qquad (5.1)$$

Eine Regression, die alle Variablen X_k beinhaltet, gibt ein für den Datensatz der Wochenenddaten korrigiertes $R^2 = 0{,}143$, für den Datensatz der Werktage lediglich ein korrigiertes $R^2 = 0{,}111$. Beide R^2 sind nahe 0, was laut Fahrmeir (2009) eine geringe Anpassung des Modells an die Daten impliziert. Beide Modelle enthalten jedoch einige Variablen, die einen Signifikanzwert t-value größer als das Signifikanzniveau von α =10% vorweisen. Im nächsten Absatz soll daher anhand der Rückwärts-Selektion nicht signifikante Regressoren aus dem Modell (5.1) ausgeschossen werden.

5.1.2 Rückwärts-Selektion

Die Methode der Rückwärts-Selektion startet zunächst mit einem Modell, welches alle Kovariablen enthält. In jedem Durchlauf werden dann schrittweise Variable aus dem Modell entfernt, bis letztendlich das Modell mit der höchsten Güte erschlossen wurde. Die Methode der Rückwärts-Selektion dient neben der Korrelationsmatrix außerdem zur Vermeidung von Multikollinearität in den Daten. Außerdem werden diejenigen Variablen entfernt, die nichts zur Erklärung der Varianz der Variable *Y* beitragen.

Die Rückwärts-Selektion führt in dem Modell der Wochenend- und Werktagsdaten dazu, dass die Variable *lorEinwohnerFlaeche* ausgeschlossen wird, was bereits in der Zusammenhangsanalyse festgestellt wurde. Außerdem werden die Variablen *lorWohnnutzung, lorÖPNV* und einzelne Variablen zur Studenfensterangabe in beiden Modellen nicht berücksichtigt. Auffällig ist, dass das Modell für den Datensatz der Werktagsdaten wesentlich mehr erklärende Variablen beinhaltet als das Modell für den Datensatz der Wochenenddaten. Sowohl die Variablen *lorGewerbenutzung, lorVerkehrsfläche* als auch die Variablen *dummyNiederschlag, Fahrzeugtyp, Akkustand* und *aktiveRoller* sind auf einem Signifikanzniveau von 10% der Werktagsdaten signifikant. Letztendlich führt die Rückwärts-Selektion jedoch mit einem korrigierten R^2= 0,142 und der Berücksichtigung von 21 Kovariablen für den Datensatz der Werktagsdaten zu keiner

verbesserten Erklärung der Varianz als das Modell, welches durch die Methode Einschluss alle Variablen mit berücksichtigt. Auch für den Datensatz der Wochenenddaten weist das letztendlich beste Modell erneut nur ein korrigiertes $R^2 = 0{,}111$ unter Berücksichtigung von 25 Kovariablen auf.

Im direkten Vergleich zu dem anfangs aufgestellten Modell (5.1), lässt sich durch die Methode einer Rückwärts-Selektion in beiden Datensätzen kein höheres korrigiertes R^2 erzielen.

5.1.3 Residuenanalyse

Zusätzlich zu der Betrachtung der korrigierten R^2, der Regressionskoeffizienten und deren Signifikanzwerte lässt sich anhand einer Residuenanalyse überprüfen, ob das Regressionsmodell die Annahmen eines linearen Regressionsmodells (s.Abschnitt 4.1.1) erfüllt. Die Residuen einer Regression geben Auskunft über die Abweichungen der beobachteten von den theoretisch zu erwartenden Werten. Diese Abweichungen sollen normalverteil sein und zufällig auftreten, es darf keine systematische Streuung erkennbar sein. Dazu werden die standardisierten Residuen gegen die standardisierten vorhergesagten Residuen in einem Streudiagramm abgebildet. Das Diagramm dient der Überprüfung der Linearität und der Varianzhomogenität.

Das Histogramm der standardisierten Residuen in Abbildung 11 zeigt eine deutliche Abweichung von der Normalverteilung. Auch das Streudiagramm in Abbildung 12 weist keine zufällige Verteilung der Residuen auf. Die Residuen streuen über den Wertebereich von *Y* sehr unterschiedlich, was auf die Verletzung der Annahme homogener Fehlvarianzen hinweist.

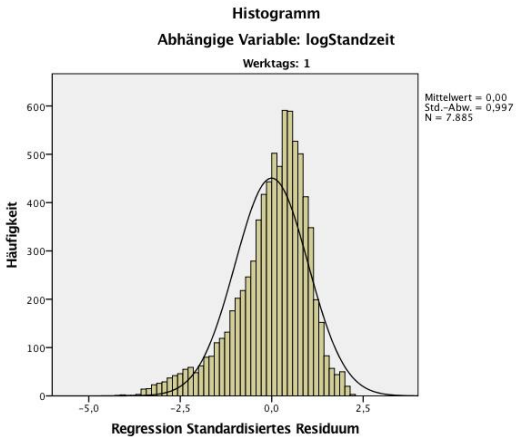

Abbildung 11: Histogramm der Residuen

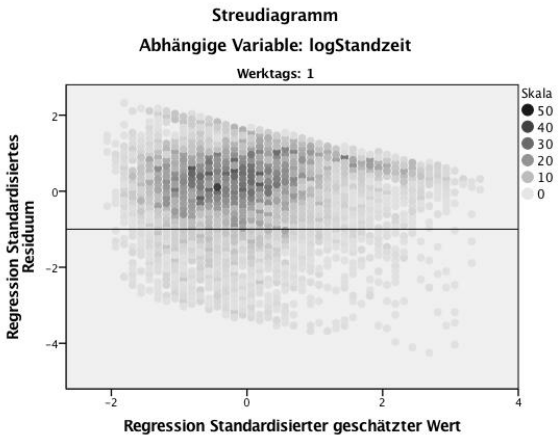

Abbildung 12: Streudiagramm der Residuen

Durch die Ergebnisse der Regressionsmodelle ist ersichtlich, dass das aufgestellte Modell (5.1) nicht gut an die vorhandenen Datensätze angepasst ist. Auch durch die Rückwärts-Selektion lässt sich kein höheres R^2 erzielen. Die Residuenanalyse liefert weitere Hinweise, dass die Variable *logStandzeit* nicht ohne weitere Datenmodifikation, Transformation oder weiterer Hinzunahme relevanter erklärender Variablen durch die Regressoren erklärt werden kann. Im folgenden Abschnitt wird daher eine weitere statistische Methode zur Aufstellung eines Regressionsmodells zur Erklärung der Standzeiten durchgeführt.

5.2 Interpretation der Ergebnisse aus der Cox-Regression

Im Gegensatz zum linearen Regressionmodell, wird im Cox-Regressionsmodell nun die untransformierte Variable *standzeit* in Sekunden als zu erklärende Variable *Y* verwendet. In der Literatur wird in diesem Zusammenhang der Begriff der Überlebenszeit bzw. der Überlebenswahrscheinlichkeit verwendet.

Die Neuanmietung eines Rollers bildet das Zielereignis, die Standzeit eines Rollers zwischen dem Abstellen bis hin zur Neuanmietung ist daher das zu untersuchende Zeitintervall.

Der Datensatz umfasst nur vollständige Ereignisse, dh. alle Werte sind gültig. Um lediglich signifikante und aussagekräftige erklärende Variablen in das Modell aufzunehmen, wird erneut schrittweise selektiert und die Variablen mithilfe der Wald-Statistik, welche die Signifikanz der berechneten Koeffizienten β_k anzeigt, ausgewählt. Im Folgenden werden die Ergebnisse erläutert und der proportionale Einfluss einzelner Kovariablen interpretiert. Die Datensätze sind erneut in Werktags- und Wochenenddaten gruppiert. Zur Berechnung des Risikos, dass ein Roller in einem bestimmten Zeitintervall *t* angemietet wird, wie folgt modelliert:

$$\begin{aligned} h(t) = h_0(t) \cdot \exp(&\beta_1 T_t + \beta_2 NS_t + \beta_3 WS_t \\ &+\beta_4 FT + \beta_5 AS + \beta_6 AR \\ &+\beta_7 WN_{lor} + \beta_8 GN_{lor} + \beta_9 VF_{lor} + \beta_{10} K_{lor} + \beta_{11} \ddot{O}PNV_{lor} \\ &+\beta_{12_{t=1}} S + \ \ldots \ + \beta_{25_{t=23}} S) \end{aligned} \qquad (5.2)$$

5.2.1 Analyse der Kovariaten

Nicht signifikante Variablen und damit aus der Regression ausgeschlossen sind *lorWohnnutzung, lorEinwohnerFläche, lorÖPNV, Wind* und einzelne Stundenvariablen. Im Datensatz der Wochenenddaten werde außerdem die Variablen *Fahrzeugtyp, Akkustand, lorGewerbenutzung* und *lorVerkehrsfläche* eliminiert. Alle übrigen Variablen werden in das Modell mit eingeschlossen und im Folgenden näher erläutert.

Je weiter der Wert der Hazartrate (als Exp(B) im SPSS-Output vermerkt) einer Kovariaten von dem Wert 1 entfernt liegt, desto stärker der Einfluss auf die Standzeit. Ist Exp(B) kleiner Null, so erhöht sich das Risiko durch die Variable; ist Exp(B) größer Null, verringert sich das Risiko einer längeren Standzeit.

Vergleicht man die negativ doppelten Werte des Logarithmus der beiden Ergebnisse miteinander, zeigt sich ein geringerer Wert von knapp 44.425 für den Datensatz der Wochenenden im Vergleich zu -2LL = 116.605 an Werktagen. Beide Werte verändern sich

durch Eliminierung einzelner Variablen nicht. Im Folgenden werden zunächst einzelne Ergebnisse der Werktagsdaten erläutert, die Ergebnisse der Wochenenddaten liefern ähnliche Werte der Hazardraten, wobei auf die genaue Auflistung der Werte verzichtet wird.

			Hazardrate
	B	**Signifikanz**	**Exp(B)**
Temperatur	,016	,000	1,016
dummyNiederschlag	-,171	,000	,842
Fahrzeugtyp		,005	
Fahrzeugtyp(1)	,008	,859	,992
Fahrzeugtyp(2)	,074	,029	1,077
Fahrzeugtyp(3)	,119	,006	1,126
lorVerkehrsfläche	-1,305	,000	,271
lorKundenFlaeche	,001	,000	1,001
Stunde_15	1,028	,000	2,794
Stunde_16	1,159	,000	3,185
Stunde_17	1,246	,000	3,475
Stunde_18	1,166	,000	2,209

Tabelle 4: Hazardrate der Kovariaten

Die Dummy-Variablen zur Startzeitstunde beziehen auf die Referenzgruppe des Stundenfensters 0. Die Variablen zum Stundenfenster 1 bis einschließlich 5 werden aus Gründen der Signifikanz in der Rückwärts-Selektion aus dem Regressionsmodell ausgeschlossen. Im Blick auf die übrigen Zeitvariablen zeigt sich ein klarer Anstieg der Hazardrate bis hin zu $S = 17$, welches eine Hazardrate $Exp(B) = 3{,}475$ aufweist. Roller, die demnach zwischen 17:00 und 17:59Uhr abgestellt wurden, haben bei gleichbleiben aller anderen Variablen ein fast 3,5-fach so hohes Risiko neu angemietet zu werden als Roller, die im Zeitraum von 00:00 und 00:59Uhr abgestellt wurden.

Der Betracht der Signifikanzen und Hazardraten der Fahrzeugtypen gibt außerdem Aufschluss darüber, dass zwischen den Fahrzeugmodellen Novi und Nova kein signifikanter Unterschied besteht. Den deutlichsten Unterschied zur Referenzgruppe des Fahrzeugtyps Novi verzeichnet der Typ Schwalbe, welcher im Jahr 2017 neu in die emmy Flotte aufgenommen wurde und die geringsten mittleren Standzeiten aufweist. Die Hazardrate $Exp(B)$1,126 lässt sich wie folgt interpretieren: Der Fahrzeugtyp Schwalbe hat ein fast 1,13-fach so hohes Risiko der erneuten Anmietung im Vergleich zum Fahrzeugtypen Novi. Die Hazardrate der stetigen Variable *Temperatur* von 0,016 sagt aus, dass wenn die Temperatur um 1°C steigt, somit das Risiko einer Neuanmietung um 0,16 % höher ist. Außerdem sinkt das Risiko einer Neuanmietung um 1,71%, wenn es während dem Stundenfenster, in dem

der Roller abgestellt wird, regnet. Ebenso hat die Variable *lorVerkehrsfläche* einen negativen Einfluss auf das Neuanmietungsrisiko. Ist der Anteil der Verkehrsfläche in einem LOR im Verhältnis zur Gesamtfläche des LOR um 1% höher, so ist das Risiko der Neuanmietung um 130,5% geringer. In allgemeiner Form kann die Hazardrate $h(t)$ in Abhängigkeit der Zeit *t* für den Werktagsdatensatz wie folgt für *i=1,…,n* bestimmt werden:

$$\begin{aligned} h(t) = h_0(t) \cdot \exp(&0{,}017 \cdot T_t - 0{,}158 \cdot NS_t - 0{,}008 \cdot FT_{Nuva} \\ &+0{,}074 \cdot FT_{Muvi} + 0{,}119 \cdot FT_{Schwalbe} - 0{,}002 \cdot AS + 0{,}519 \cdot AR \\ &-0{,}357 \cdot GN_{lor} - 1{,}305 \cdot VF_{lor} + 0{,}001 \cdot K_{lor} \\ &+0{,}370_{t=6} \cdot S + \ \ldots + 0{,}220_{t=23} \cdot S) \end{aligned} \qquad (5.3)$$

5.2.2 Überlebensfunktion

Die zum Mittelwert der Kovariaten gehörenden Überlebensfunktion ist in Abbildung 13 grafisch dargestellt. Die Wahrscheinlichkeit, dass die Zeit bis zur Neuanmietung 10.000 Sekunden Standzeit (ca. 2,7 Stunden) beträgt, lässt sich anhand der Grafik an der y-Achse bei ca. 30% ablesen. Der Median der Überlebenszeit, d.h. 50% der Roller sollten nach diesem Zeitpunkt bereits neu angemietet sein, lässt sich anhand der x-Achse in der Grafik bei ca. 5.000 Sekunden Standzeit (ca. 1,40 Stunden) ablesen. Voraussetzung für die Annahme der Überlebensfunktion ist, dass die Kovariaten exakt den Wert ihres Mittelwerts betragen. Die rechte Grafik von Abbildung 13 zeigt die unterschiedlichen Verläufe der Überlebensfunktion bzw. des Neuanmietungsrisikos im Vergleich des Stundenfensters 17 zum Rest des Tages an Werktagen. Wird ein Roller in der Zeit zwischen 17:00 und 17:59Uhr abgestellt, ist das Risiko bei weniger als 20%, dass der Roller eine Standzeit von 1,40Stunden aufweist.

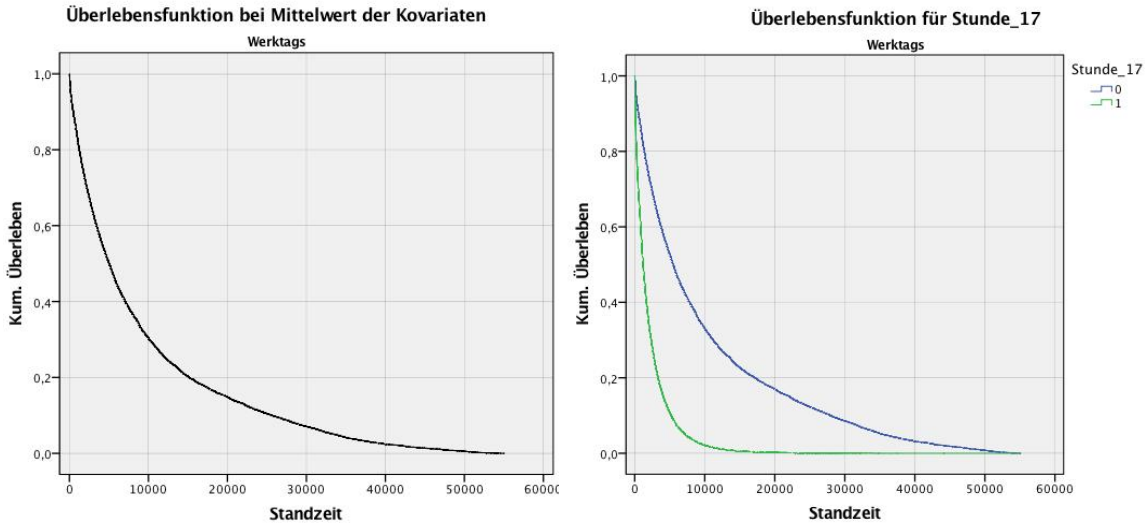

Abbildung 13: Überlebensfunktion bei Mittelwert der Kovariaten

6. Fazit und Ausblick

Ziel dieser Arbeit war es, ein Modell aufzustellen, welches die Standzeiten der Elektroroller in einem FFSS erklärt. Es wurden Daten aus verschiedenen Quellen zusammengefügt, um im Anschluss die Nutzerdaten des Unternehmens emmy analysieren zu können.
Die räumliche Darstellung der Standzeiten im Berliner Geschäftsgebiet weist auf, dass gerade die Gebiete im Prenzlauer Berg, Kreuzberg und Friedrichshain im Vergleich zu den restlichen Gebieten die kürzesten Standzeiten aufzeigen. Auch die zeitliche Variable hat einen Einfluss auf die Rollernachfrage, so weist vor allem das Gebiet in Mitte und um den Bahnhof Zoologischer Garten während des Feierabendverkehrs im Durchschnitt kurze Standzeiten auf.
Als signifikant und besonders bedeutend im linearen Regressionsmodell haben sich die Variablen zur Temperatur in °C und die Anzahl der registrierten Kunden in einem LOR im Verhältnis zur LOR-Fläche herausgestellt. Jedoch weisen alle durchgeführten linearen Regressionsmodelle, sowohl für die Daten an den Werktagen als auch für die Daten am Wochenende, insgesamt keinen hohen Erklärungswert der Standzeiten auf.
Neben dem linearen Regressionsmodell dient die Cox-Regression als zweites Modell, welches ähnliche Variablen selektiert. Vor allem die Variablen zum Zeitfenster, in dem ein Roller zur Neuanmietung zur Verfügung gestellt wird, weist hier einen großen Einfluss auf das Neuanmietungsrisiko auf. Auch der Fahrzeugtyp hat bei der Wahl, welcher Roller von den Kunden angemietet wird, einen signifikanten Einfluss. Der Akkustand des reservierten Rollers hingegen scheint kein wichtiger Faktor bei der Rollerreservierung zu sein. Im Vergleich zu den linearen Regressionsmodellen, welche lediglich eine maximale Erklärung der Varianz von knapp 15% erzielten, gibt die Zielvariable des Risikos im aufgestellten Cox-Regressionsmodell eine bessere Grundlage, um auf diesem Modell aufbauend eine Relokalisierungsstrategie zu entwickeln.

Trotz alledem ist aus den insgesamt schwachen Zusammenhängen der erklärenden Variablen mit der Variable Standzeit zu schließen, dass die vorliegenden Daten nicht ausreichen, um die Standzeit eines Rollers mittels eines Regressionsmodells ausreichend zu erklären. Die Residualanalyse in Abschnitt 5.1.3 gibt außerdem darüber Auskunft, dass im Datensatz ein Muster enthalten ist, was durch die bisherige Datenanalyse nicht eliminiert wurde, weswegen eine weitere Transformation der Daten nützlich sein könnte. Vergleicht man die Ergebnisse dieser Arbeit mit anderen Forschungen des Car-, Roller- oder Bikesharing wird deutlich, dass die deskriptive Statistik aller Forschungsansätze eindeutige Ergebnisse liefert. Die Zeiten und Orte einer jeder Stadt, die am häufigsten nachgefragt werden, sind eindeutig

aus den vorhandenen Daten herauszufiltern. Jedoch liefert keiner der Ansätze eine allgemeine Formel, anhand derer die Nachfrage in Abhängigkeit von verschiedener Faktoren abgebildet werden kann. Weitere Erkenntnisse könnten durch zusätzliche Datenerfassung gewonnen werden, wie beispielsweise die Erfassung der Kundenabsicht, einen Roller an einem bestimmten Ort zu einer bestimmten Zeit zu mieten, oder die Klassifizierung der Kunden in Abhängigkeit zur Häufigkeit, in der jeder Kunde einen Sharinganbieter nutzt. Schmöller et al. schlagen in ihrer Arbeit aus dem Jahr 2015 bereits vor, dass eine Umfrage der Kunden weitere Aufschlüsse erbringen könnte und so die Nutzerdaten nachvollziehbarer wären. Die Standzeitenanalyse in dieser Arbeit beinhaltet interessante Erkenntnisse und bildet eine umfangreiche Grundlage für weitere Erforschungen und Erkenntnisse über das Nutzerverhalten auf dem Markt von FFSS.

Literaturverzeichnis

Berliner Morgenpost (2017): So tickt Berlin an deiner Linie [online]: https://interaktiv.morgenpost.de/berlin-an-deiner-linie/, Abrufdatum: 05.12.17.

Bogenberger, K., Schmöller, S. (2014): Analyzing External Factors on the Spatial and Temporal Demand of Car Sharing System, in: Procedia-Social and Behavioral Sciences 111: S.8-17.

Bogenberger, K., Müller, J. (2015): Time Series Analysis of Booking Data of A Free-Floating Carsharing System In Berlin, in: Transportation Research Procedia 10: S. 345-354.

Bogenberger, K., Müller, J. Schmöller, S., Weikl, S., (2015): Empirical analysis of free-floating carsharing usage: The Munich and Berlin case, in: Transportation Research Part C.

Bogenberger, K., Reiss, S. (2016): A Relocation Strategy for Munich's Bike Sharing System: Combining an operator-based and a user-based Scheme Combining an operator-based and a user-based Scheme, in: Transportation Research Procedia No.22: S.105 – 114.

Bundesverband CarSharing (2017): Unterschiede free-floating & stationsbasiertes CarSharing [online]: https://carsharing.de/presse/fotos/zahlen-daten/unterschiede-free-floating-stationsbasiertes-carsharing, Abrufdatum: 05.12.2017.

Carsharing Magazin (2017). Carsharing in Berlin - Carsharing Magazin. [online]: *http://www.carsharing-magazin.de/carsharing-berlin*, Abrufdatum: 05.12.2017.

Dwd.de (2017). Wetter und Klima - Deutscher Wetterdienst - Leistungen - Klimadaten Deutschland. [online]: https://www.dwd.de/DE/leistungen/klimadatendeutschland/klimadatendeutschland.html, Abrufdatum: 05.12.2017.

Emmy-sharing.de (2017). emmy | Elektro-Roller Sharing in Berlin. [online]: https://emmy-sharing.de, Abrufdatum: 05.12.2017.

Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). Multivariate Statistische Verfahren (2. Auflage), De Gruyter, Berlin.

Fahrmeir, L., Kneib, T., Lang, S. (2009). Regression: Modelle, Methoden und Anwendungen. Springer Verlag, Berlin Heidelberg.

Innoz.de (2017): Global Scootersharing Market Report. [online]: *https://www.innoz.de/sites/default/files/howebock_global_scootersharing_market_report_2017.pdf,* Abrufdatum: 05.12.2017.

Kortum, K., Machemehl, R. (2012): Free-Floating Carsharing Systems: Innovations in membership prediction, mode share, and vehicle allocation optimization methodologies. *https://repositories.lib.utexas.edu/bitstream/handle/2152/ETD-UT-2012-05-318/KORTUM-DISSERTATION.pdf?sequence=1&isAllowed=y*, Abrufdatum: 05.12.17.

Stadtentwicklung.berlin.de (2017). Lebensweltlich orientierte Räume (LOR) / Land Berlin. [online]: *http://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/lor/index.shtml*, Abrufdatum: 12.12.17.

Anhang A

Kollinearitätsstatistik

	Toleranz	VIF
Temperatur	,524	1,907
dummyNiederschlag	,904	1,106
Wind	,811	1,233
Fahrzeugtyp	,991	1,009
Akkustand	,974	1,027
aktiveRoller	,914	1,094
lorWohnnutzung	,163	6,134
lorGewerbenutzung	,851	1,175
lorVerkehrsfläche	,744	1,343
lorEinwohnerFlache	,075	13,256
lorKundenFlaeche	,185	5,403
lorÖPNV	,781	1,280
Stunde_01	,558	1,792
Stunde_02	,666	1,502
Stunde_03	,741	1,349
Stunde_04	,762	1,313
Stunde_05	,857	1,167
Stunde_06	,854	1,171
Stunde_07	,780	1,282
Stunde_08	,592	1,689
Stunde_09	,499	2,003
Stunde_10	,424	2,358
Stunde_11	,391	2,554
Stunde_12	,387	2,584
Stunde_13	,372	2,689
Stunde_14	,345	2,894
Stunde_15	,341	2,937
Stunde_16	,330	3,033
Stunde_17	,339	2,952
Stunde_18	,335	2,989
Stunde_19	,356	2,811
Stunde_20	,335	2,986
Stunde_21	,375	2,666
Stunde_22	,398	2,512
Stunde_23	,468	2,138

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides Statt, dass ich die vorliegende Arbeit mit dem Titel „Standzeitenanalyse von Elektrorollern im Free-Floating-Sharing-System in Berlin“ selbstständig und nur unter Zuhilfenahme der ausgewiesenen Hilfsmittel angefertigt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach anderen gedruckten oder im Internet verfügbaren Werken entnommen sind, habe ich durch genaue Quellenangaben kenntlich gemacht.

_____

Ort, Datum

_____ _____

Name, Vorname Unterschrift